u s e n i x
ASSOCIATION

Twentieth Symposium on Usable Privacy and Security (SOUPS 2024)

Philadelphia, PA, USA
August 12–13, 2024

Sponsored by

u s e n i x
THE ADVANCED
COMPUTING SYSTEMS
ASSOCIATION

SOUPS 2024 Sponsors

Gold Sponsors

Silver Sponsor

Bronze Sponsor

USENIX Supporters

USENIX Patrons

Futurewei • Google • Meta

USENIX Benefactor

Bloomberg

USENIX Partner

Thinkst Canary

Open Access Supporter

Google

Open Access Publishing Partner

PeerJ

USENIX Association

Proceedings of the
Twentieth Symposium on Usable Privacy
and Security (SOUPS 2024)

August 12–13, 2024
Philadelphia, PA, USA

Symposium Organizers

External Reviewers

Message from the
SOUPS 2024 Program Co-Chairs

Welcome to SOUPS 2024!

This year we proudly celebrate two decades of SOUPS! For its 20th anniversary, our SOUPS community has collectively ensured another excellent and exciting conference program, this year with bonus celebratory activities. With 33 papers accepted out of 156 submissions (21.2% acceptance rate), the technical program covers a wide range of topics within usable privacy and security. The conference also includes workshops, posters, lightning talks, mentorship activities, networking, celebrations, and a 20th anniversary panel.

In 2016, SOUPS became an independent conference body. Since then, we have partnered with USENIX for hosting and administrative support, a move that has enabled continued growth and stability for the conference. We thank all the members of the USENIX staff for their work in organizing SOUPS and supporting our community. Their team has been fantastic at making the process of running a high-quality conference seamless.

In 2018, we co-located with the USENIX Security Symposium for the first time, and we have continued that co-location for 2024. Co-locating the two conferences in-person allows for interactions and shared ideas between SOUPS and USENIX Security attendees. We have found this beneficial for both conferences and look forward to the opportunity again this year as USENIX welcomes back other co-hosted events in parallel to SOUPS.

SOUPS relies on a range of volunteers for all of its activities. Steering Committee members provide oversight and guidance and are elected for three-year terms. Organizing Committee members help determine the conference content for a particular year, often serving two-year terms to facilitate the transition of knowledge. Technical Papers Committee members are chosen by the Technical Papers Co-Chairs each year. SOUPS is a product of the hard work of many people, starting with researchers who decide to submit their work to SOUPS, and including all of the SOUPS Organizers, the SOUPS Steering Committee, the technical paper reviewers, the workshop organizers, the poster jury, and the USENIX staff. We are grateful and thank each and every one of you for your contributions to SOUPS 2024.

Apu Kapadia has served as General Co-Chair of SOUPS since 2022 and Chair of the Steering Committee since 2023. He is ending his term with this iteration of SOUPS (2024). Patrick Gage Kelley was appointed as Vice Chair in 2023 and is General Co-Chair for SOUPS in 2024. If you are interested in helping with SOUPS 2025 in Seattle (Aug 10–12, 2025), in any capacity, please contact General Chair Patrick Gage Kelley.

SOUPS would not be possible without the generous support of our sponsors–thank you for supporting this community this year, and over the past two decades.

Please visit our website to view the recipients of the SOUPS 2024 awards. Congratulations to all recipients for their outstanding work.

And, once again, Happy Twentieth Anniversary to SOUPS!

Patrick Gage Kelley, *Google, General Co-Chair*
Apu Kapadia, *Indiana University, General Co-Chair*
Katharina Krombholz, *CISPA Helmholtz Center for Information Security, Technical Papers Co-Chair*
Mainack Mondal, *Indian Institute of Technology Kharagpur, Technical Papers Co-Chair*

Twentieth Symposium
on Usable Privacy and Security (SOUPS 2024)

August 12–13, 2024
Philadelphia, PA, USA

Monday, August 12

Expert Community

Tuesday, August 13

Online Community

Mobile Security

Security in the Workplace

Social Aspects of Security

A Survey of Cybersecurity Professionals' Perceptions and Experiences of Safety and Belonging in the Community

Samantha Katcher[*▷], Liana Wang[*], Caroline Yang[*], Chloé Messdaghi[‡],
Michelle L. Mazurek[†], Marshini Chetty[◇], Kelsey R. Fulton[∪], and Daniel Votipka[*]
[*]*Tufts University;* [‡]*SustainCyber;* [†]*University of Maryland;*
[◇]*University of Chicago;* [∪]*Colorado School of Mines;* [▷]*MITRE Corporation*

Abstract

The cybersecurity workforce lacks diversity; the field is predominately men and White or Asian, with only 10% identifying as women, Latine, or Black. Previous studies identified access to supportive communities as a possible disparity between marginalized and non-marginalized cybersecurity professional populations and highlighted this support as a key to career success. We focus on these community experiences by conducting a survey of 342 cybersecurity professionals to identify differences in perceptions and experiences of belonging across demographic groups. Our results show a discrepancy between experiences for different gender identities, with women being more likely than men to report instances of harassment and encountering unsupportive environments because of their gender. Psychological safety was low across all demographic groups, meaning participants did not feel comfortable engaging with or speaking up in the community. Based on these result we provide recommendations to community leaders.

1 Introduction

With technology's growing ubiquity, and parallel increases in cyberattacks, skilled cybersecurity professionals are in demand. This demand has outpaced the supply of qualified workers, with some estimates suggesting a four million job shortfall in 2023 [54]. Governments and private institutions are campaigning to increase the number of cybersecurity professionals [8, 35, 36, 51, 85, 110, 113] and the US government has prioritized growing the cybersecurity workforce [7].

While there have been many efforts to grow the cybersecurity workforce, this growth has not increased diversity. Prior workforce surveys show the field as predominantly male, white or Asian[1], with women, Latine, and Black participants constituting fewer than 10% in each survey [10, 45]. In a 2020 cybersecurity professionals survey, SynAck, a platform connecting organizations with cybersecurity professionals who provide security reviews, found women (66%) and members of marginalized ethnicities (47%) were less likely, when compared to men (88%), to believe people of the same gender or ethnicity were given the same opportunities [102]. Furthermore, interviews with cybersecurity professionals from marginalized populations revealed regular instances of othering, hate, and harassment in the workforce [38].

Cybersecurity's deficiency in diversity creates two problems. First, and foremost, is an equity problem. Members of marginalized populations are driven away from well paying, in-demand careers in cybersecurity. Second, cognitive diversity is essential to secure system design. The more eyes reviewing potentially insecure code, the more thorough a review will be completed and attacks thwarted [63]. People from different genders, ethnicities, and backgrounds provide a fresh perspective to solving complex security problems [28, 69]. As cybersecurity hiring increases, we must prevent furthering existing ethnic and gender disparities by identifying and understanding factors underlying lacking diversity to improve recruitment and retention among marginalized populations.

Recent work investigates the career challenges cybersecurity professionals face through a survey broadly with cybersecurity professionals [2] and interviews with marginalized cybersecurity professionals [38]. Both populations indicated that the most significant challenges were the result of the difficulty of getting started, e.g., navigating unstructured resources to develop necessary skills, and the stress and uncertainty of the market, e.g., trying to find work for which they qualified. Non-marginalized cybersecurity professionals found support

---

[1]We use the term Asian broadly here, as this is how it is used in the cited prior surveys, but we recognize Asian Americans and those from other regions may still be marginalized in the community.

from their peers crucial, viewing the community as inclusive, while marginalized cybersecurity professionals found it challenging or impossible to join the community, hindering their access to necessary support for success.

In this paper, we take further steps to understand the barriers faced by marginalized cybersecurity professionals by focusing on this point of divergence: their community experiences. We surveyed 342 cybersecurity professionals from varying backgrounds (196 men, 128 women, 10 genderqueer; 215 White, 46 Latine, 38 Black, and 31 other ethnicities). We used multiple validated psychometric scales to measure perceptions of belonging [30, 122] and experiences of supportive and unsupportive social environments [43, 108]. We also asked about participation and experiences in specific subcommunities. We address the following research questions:

- **RQ1:** What differences exist in perception (e.g., belonging, psychological safety) and incidents of unsupportive experiences (e.g., othering, hate, harassment) within the cybersecurity community between marginalized and non-marginalized cybersecurity professionals?

- **RQ2:** Do marginalized and non-marginalized cybersecurity professionals differ in their participation and experiences in specific subcommunities (e.g., work, social organizations, online)?

- **RQ3:** What community interactions are perceived as particularly supportive or unsupportive and how do these differ between marginalized and non-marginalized cybersecurity professionals?

The biggest divide among cybersecurity professionals was across gender identities, with women being more likely to report experiencing harassment and unsupportive environments due to their identity. However, across all demographic groups, cybersecurity professionals reported low psychological safety relative to other professions, indicating the difficultly to engage in the community. Conversely, we did not observe low scores on measures of internal belonging (i.e., whether a participant felt qualified and knowledgable enough to belong in the community). Together these suggest unsupportive forces on cybersecurity professionals are generally external to the individual. Finally, our results suggest early development environments for cybersecurity professionals might be particularly problematic since participants with high school programming experience were less likely to feel psychologically safe. Based on our results, we provide recommendations for cybersecurity community leaders.

2   Related Work

Our study's contribution lies in a focused exploration of belonging in the cybersecurity community, differentiating it from other studies [38, 91, 106, 123]. Here we describe prior work and how our study fits into the broader research context.

Marginalized populations' experiences in computer science and technology. There is a growing body of research considering issues facing marginalized populations in CS and STEM domains. For example, work studying developers has found marginalized populations are paid less [42, 73] and are less likely to have work accepted by colleagues [9, 74, 105, 116]. Similarly, significant research has investigated issues in CS [16, 17, 18, 34, 64, 65, 92, 95, 123] and technology careers more broadly [14, 86, 124]. Margolis and Allen performed an ethnographic study of the gender gap in CS education [65]. They found women had less coding experience than men in undergraduate programs and perceived CS's "geek" culture negatively. Subsequent work has documented issues of gendered perceptions of CS [15, 22, 72, 76, 93], which are further entrenched by unapproachable early educational activities [3, 19, 60] , lack of representation [3, 109], mentorship support [1, 3, 19, 109, 126], and a non-inclusive culture to diverse backgrounds and experiences [1]. Because we expect many of these trends to be mirrored in cybersecurity, we use this prior work as a lens, guiding our survey questions and analysis. However, we expect cybersecurity may present differences as it is more specialized and the inherent focus on privacy and security scrutiny may make cybersecurity communities less welcoming. This has been found, to some extent, demonstrating several differences in interviewee experiences when studying members of the vulnerability discovery community—a subset of the cybersecurity community [38].

Marginalized populations' experiences in cybersecurity. Several prior industry surveys have demonstrated the lack of diversity in the cybersecurity community [10, 44, 54, 102]. This includes ICS2's annual survey of the cybersecurity workforce, which showed that the younger generation (under 30 years old) of cybersecurity professionals are more diverse. However, this diversity remains limited as only 26% of this generation are women [54]. This survey also found the pathways into cybersecurity differ by gender and race/ethnicity. Women and non-white cybersecurity professionals are more likely to come from a traditional education-based pathway (e.g. college) and less likely to come from an IT background.

In addition to these industry surveys, some academic interview-based studies have examined the challenges marginalized cybersecurity professionals face [38, 83, 91, 106, 112]. Fulton et al. conducted semi-structured interviews with members of the vulnerability discovery community from marginalized populations, uncovering challenges specific to members of marginalized populations, such as a difficulty being taken seriously by others, a reluctance from other community members to share information, and explicit discrimination within the community. Additionally, Fulton et al's works discussed the important role mentors played in participants' experiences [38]. In interviews with 21 cybersecurity

professionals, Schoenmaker et al. found some participants believed holding a minority status might cause an increase in an individual's ability to monitor for security anomalies, as these individuals already have significant experience monitoring for threats regarding personal safety. However, they also observed social conventions and lack of access to resources might make it more difficult for these groups to practice vulnerability discovery [91]. Plato et al. interviewed sixteen women C-Suite executives in cybersecurity to learn about their journeys into leadership and experiences with mentorship, sponsorship, and trusted advisors, as well as experiences of biases and discrimination highlighting how networking, mentorship, and observing leadership styles play pivotal roles in shaping individuals' trajectories, even with a shortage of female mentors and racial bias making this difficult to accomplish in practice [83]. Each of these studies highlights important challenges marginalized cybersecurity professionals face, but have limited generalizability due to their small samples. Our work expands on these findings with a large-scale survey focusing on community belonging, a central challenge observed in prior work.

Students' cybersecurity experiences. Some work has investigated existing workforce disparities. This work has primarily focused on student experiences in security exercises [29, 84] and college courses [13] as students take the first steps toward cybersecurity careers. It provides some indications of students' reasons for abandoning the field (e.g., lack of role models and community, gendered stereotypes) and suggests entry-level hands-on exercises can increase interest. While these education-focused questions are important, cybersecurity professionals face challenges throughout their careers [38] and prior work has found many are not trained through these traditional educational settings [45, 121]. To address this gap, our work takes a holistic view of cybersecurity professionals' community environments.

3 Methods

We seek to understand how practitioners in cybersecurity participate in and perceive belonging within their professional community, and specifically to consider differences and similarities between practitioners from different demographics. We do not place limits on participation (e.g., industry, academia, government), but consider the field of cybersecurity broadly. In this section, we describe the survey design, recruitment methods, data analysis procedures, ethical considerations, and the work's limitations.

3.1 Survey Design

The survey began by requesting participant consent; included three main components aligned with the research questions;

and concluded with demographics questions. Figure 1 summarizes the survey's flow. The full survey can be found in Appendix A. Where applicable, we altered validated scales to focus on cybersecurity and the cybersecurity community, and we included attention checks to catch inattentive respondents [68]. The survey was divided into three parts to match our research questions: questions about participants' belonging within the cybersecurity community generally (RQ1), participation in various subcommunities–listed in Table 2–(RQ2), and prototypical community experiences (RQ3). We detail how we asked about each of this topics in turn, then concluded with questions about their security experience and demographics. Figure 1 summarizes the survey's flow. Participants completed the survey in 15 minutes on average.

Perceptions of belonging (RQ1, Figure 1.B). We first sought to understand whether participants feel they belong in the cybersecurity community, as prior work found cybersecurity professionals were more successful after finding a community where they could get support and ask questions [38].

We utilized three validated psychometric measures of belonging: *psychological safety* [30], *belonging uncertainty* [122] and *vulnerability discovery self-efficacy* [119]. Table 1 provides additional details about each scale.

The psychological safety scale has previously been used to investigate why employees feel comfortable sharing information [21, 94], suggesting organizational improvements [27, 62], and taking initiative [5]. The belonging uncertainty scale has predominantly been used investigate feelings of otherness among historically underrepresented groups, for example, among professionals [122] and students [26]. We also ask participants explicitly whether people with similar backgrounds have opportunities to participate in cybersecurity work to assess the question of representation more directly.

Finally, to assess whether participants believed they had the skill to be in the community (i.e., separate from whether they believed others would accept them into the community) we used the vulnerability discovery sub-scale of Votipka et al.'s secure software development self-efficacy scale (SSD-SES), which asks participants to assess their proficiency to identify vulnerabilities [119]. SSD-SES has been used to assess differences in perceived ability between study subgroups [56, 103, 104] and as a measure of learning improvement with educational interventions [39, 120].

(Un)welcoming Community Experiences (RQ1, Figure 1.B). To understand concrete experiences that might impact cybersecurity professionals' community participation, we asked a modified version of de Grey et al.'s Online Social Experiences Measure (OSEM), which assesses social support and negativity arising from online social network interactions [43]. OSEM evaluates aspects such as emotional, informational, and instrumental support. This measure has been employed in research to understand how online interactions influence mental health and social well-being, particularly in
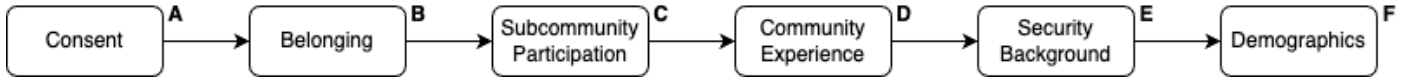
Figure 1: Structure of the survey.

Name	Description	Ex. Item	Response Opts.	Items	Agg[1]
Psychological Safety [30]	whether participants feel safe to express themselves, take risks, and ask questions	"If you make a mistake in the cybersecurity community, it is often held against you"	7-pt; "Very inaccurate" to "Very accurate"	5	Avg
Belonging Uncertainty [122]	Whether participants feel people like them belong in the community	"When something bad happens, I feel that maybe I don't belong in cybersecurity"	7-pt; "Strongly disagree" to "Strongly agree"	3	Avg
Vuln Discovery Self-Efficacy [119]	Whether participants believe they have appropriate cybersecurity skills	"I can identify potential security threats to the system"	5-pt; "Not confident at all" to Absolutely confident"	9	Sum
Online Social Experiences [43]	How often participants experience positive and negative interactions	"Someone in the cybersecurity community has made me feel embarrassed or foolish"	5-pt; "Very Slightly or Not at All" to "Extremely"	8	Sum
Hate and Harassment [108, 117]	How often participants experience hate and harassment	"Stereotyping based on perceived demographic characteristics"	4-pt; "Never" to "Frequently"	7	Sum

[1]Aggregation function used to combine responses from multiple items to a single score.

Table 1: Summary of psychometric measures used in the survey to understand participants' sense of belonging and experiences in the cybersecurity community. The different scales were presented in a randomized order to avoid ordering effects.

digital communities [20, 75, 111].

To capture hate and harassment, we borrowed from Thomas et al.'s [108] and the Pew Research Center's [117] existing survey questions investigating online hate and harassment. We included four questions about severe negative experiences which OSEM did not include, namely stereotype bias, violence, sexual advances, and doxxing. These questions have been employed in various research contexts to measure experiences of sexual harassment, particularly in professional and educational settings [24, 57, 81].

Subcommunity participation (RQ2, Figure 1.C). Community is not a global construct, but instead is specific to the individual [90]. Someone might not feel comfortable communicating with others at a large security conference, but may establish a smaller local community where they feel strong connections and receive support. Therefore, we investigated how participants' experiences varied across subcommunities–specifically, those descried in Table 2, drawn from prior work [38]. We asked how many of each type of subcommunity participants were members of, how frequently they discussed cybersecurity concepts and how helpful they found each subcommunity, and, for each subcommunity, we asked at least one subcommunity-specific question to allow a better understanding of participant's relationship to the subcommunity. We randomized the order participants were asked about each subcommunity to avoid ordering effects.

Examples of supportive and unsupportive community ex-

periences (RQ3, Figure 1.D). Next, we sought to understand what makes participants feel particularly welcome or unwelcome. We asked participants to describe an experience where they felt particularly well supported which could involve explicit assistance, encouragement, or any positive influence that aided the participant's professional growth. We also asked participants to describe a particularly unsupportive experience, which could include instances where the interaction was harmful or hindered their professional progress.

Cybersecurity background and demographics (Figure 1.D/E. We finished by asking about participants' cybersecurity background, i.e., the extent their work focuses on cybersecurity, whether and what kind of cybersecurity training they have received, when they began programming, the age they became interested in cybersecurity, and the age they first received cybersecurity career support. We ended with demographics questions like gender, ethnicity, and education.

3.2 Recruitment

Recruiting cybersecurity professionals is a difficult task because they are a small, well paid population with significant demands on their time [47, 59]. Our challenge was compounded by the fact that we weighted our sample toward marginalized cybersecurity professionals, who make up a small fraction of this small workforce [10, 45].

We used several recruit methods, including contacting cybersecurity professional organizations' leaders; advertising in

Subcommunity	Description
Close Friends / Mentors	Family, friends and other close mentors who provided either career or other support (e.g., emotional, economical, etc.)
School	A learning community focused on cybersecurity in an academic setting (e.g., class, student-run organization)
Work	Community in participants' workplaces where they are able to discuss cybersecurity topics and receive support
Organizations	Groups outside work and school (e.g., ACM chapters, Women in Security and Privacy)
Online and Conferences	The broader cybersecurity community where participants might meet and talk with other, but not have close or lasting relationships. This includes interactions at cybersecurity conferences or workshops, as well as through online forums (e.g., StackOverflow, Reddit, X (formerly Twitter), public Slack or Discord).

Table 2: Types of subcommunities participants were asked about.

public (i.e., X (formerly Twitter), LinkedIn, and Reddit) and private (i.e., Slack and Whatsapp) online spaces; recruiting at cybersecurity conferences; and contracting Qualtrics for a curated panel. Participants recruited through organizations, online, and conferences were given a study description and entered into a raffle for one of 25 $50 Amazon gift cards. For the Qualtrics panel, we instructed Qualtrics to identify paenlists working in cybersecurity, with a majority being women and at least 30% non-white. Panelists were paid $25.

Our recruitment messages indicated that anyone currently working (or having worked in the last two years) in cybersecurity could participate. We did not mention the study's intent to compare responses between marginalized and nonmarginalized cybersecurity professionals to avoid a potential backlash [55, 88] due to increasing antagonism and polarization around diversity, equity, and inclusion efforts from a segment of the population [32, 70].

3.3 Data Analysis

Next, we outline our quantitative and qualitative analyses.

Qualitative analysis. To analyze the two free response questions in Part C, we used iterative open coding [100]. Two researchers collaboratively reviewed the first 35 responses to generate the codebook. Then, the same two researchers iteratively coded responses in rounds of twenty. After each round, the coders compared responses, resolved disagreements, and updated the codebook as necessary. After six rounds of independent coding (i.e., 120 responses), the coders achieved a Krippendorff's α of 0.858 for *what* participants experienced and 0.835 for *who* the experience was with. Both are above the recommended level of agreement [46]. The remaining

Factor	Description	Baseline
Required		
Gender	Gender participants identify as	Man
Ethnicity	Ethnicity participants identify as	White
Optional		
Yrs. Exp.	Number of years participants have worked in cybersecurity	–
Yrs. Until Mentor	Age when participant first had a mentor who helped them learn about cybersecurity	–
Helpful Mentor	Whether participant reported having someone close (mentor/family member/friend) who helped them learn about cybersecurity	False
HS Prog.	Whether participant had high school programming experience	False
Job/ Seniority	Current job role (junior, senior non-leadership, senior leadership, or not currently working in cybersecurity)	Junior

Table 3: Factors used in regression models. Categorical variables are compared individually to the baseline.

responses were divided evenly among the coders and coded independently. The final codebook is included in the supplemental materials [101].

Quantitative analysis. In our statistical tests, we limited our dataset to participants who identified as men or women and were White, Black, and/or Latine. We did not include other demographics for statistical tests because there was an insufficient number of participants to produce generalizable results. We include 289 participants in the reported statistical anlaysis.

For the vulnerability discovery self-efficacy, online social experiences, and harassment questions, we used a poisson regression as the scales were scored using a sum of the Likert responses. As the harassment questions from Thomas et al. are not part of a validated scale, we first computed Cronbach's α over participants' responses to the four harassment questions to test their internal consistency [67]. These questions had "good" internal consistency ($\alpha = 0.806$), so we chose to treat them as a single measure like the other scales. For the psychological safety and belonging uncertainty scales, we used linear regressions as the outcome variables were a percentage and an average, respectively. To generate our initial models, we included all the factors listed in Table 3. Because it is possible some explanatory variables are not independent, which would violate the regressions' assumptions [12, pg. 67-106], we tested for multicollinearity and found there was no significant correlation between factors. We then conducted model selection on all possible combinations of these factors, only considering models that included gender and ethnicity as they were our key variables of interest and selected the model with minimum Bayesian Information Criteria (BIC) [87, 96].

For each subcommunity, we used Kruskal-Wallis H tests,

to compare subcommunity membership, discussion participation rates, and helpfulness Likert responses. We began each comparison with an omnibus test over all demographic groups. If the result was statistically significant, we applied planned pairwise comparisons between non-marginalized and marginalized groups for gender and ethnicity, i.e., men to women, White to Black, White to Latine.

Finally, we applied Pearson's χ^2 tests to compare responses between top-level code categories between men and women for themes that were mentioned by at least five participants [37] for our free-response questions. We focused on gender differences as we do not have sufficient data points across races/ethnicities to produce generalizable results. For categories mentioned by five or fewer of a single gender group, we perform a Fisher's Exact Test instead of a χ^2 test [33].

3.4   Ethical Considerations

This study was approved by our institution's ethical review board. Since this survey asks about multiple sensitive topics including experiences of harassment, psychological safety, and social experiences, participants were informed about our data collection and secure storage practices and that they could stop participating or skip a question at any time.

3.5   Limitations

Our study has several limitations that should be considered when interpreting our results.

Self-report responses. As is common in online studies with self-reported data, some participants may not approach the survey seriously, may try to take the survey multiple times, or may fabricate responses to qualify for compensation. To account for these behaviors, we deterred multiple responses by using a browser cookie and followed best practices for removing inattentive responses, e.g., we removed those that failed attention checks, were significantly faster than average, or provided nonsensical responses to open-ended questions (N=385). Also, we received over 500 automated responses where more than 50 identical or nearly identical responses were submitted within a very short period–often within a minute. We removed these responses as they were received as they clearly did not represent a legitimate response.

Inauthentic responses were a challenge in this study. To mitigate this, we primarily recruited from venues with a high likelihood of cybersecurity professionals, leveraging community relationships and in-person recruiting. Qualtrics panel participants were recruited independently of our study, reducing their motivation to lie, and their open-ended responses were consistent with those from professional venues.

Demographic distribution and US-centric population. While we made significant efforts to recruit participants from marginalized populations, many identities have limited representation in our sample (e.g., genderqueer, Middle Eastern, indigenous peoples). Therefore, any results from their responses may not generalize beyond our sample. We give descriptive statistics regarding their responses to provide indications of potential trends for future work to investigate, but avoid conducting statistical tests relating to these identities or making broader statements about their responses. Similarly, our small sample sizes preclude investigation into the effects of intersectional identities. We expect there are important differences introduced by intersectionality, as prior work has shown in other domains [78, 97], but we refrain from investigating them to avoid overgeneralizing their personal experiences.

Despite attempting to recruit broadly, our participant pool was predominantly US-centric, with 279 out of 342 respondents from the US. We expect experiences of cybersecurity professionals in other regions will be similar, but there are likely critically important differences; we encourage further work focus on other geographic areas.

Survivor and recall biases. Our recruitment was limited to currently or recently employed cybersecurity professionals. It is likely many members of marginalized populations considered becoming cybersecurity professionals or worked in the field, but faced substantial challenges and chose to switch professions. Our results inherently do not account for these individuals, so our findings may skew toward a more positive portrayal of cybersecuirty. We attempt to capture some of this adversity by asking participants to consider their experiences throughout their career when answering all questions, but they may not clearly remember events from years ago [89].

Demand effects. Participants might be motivated to report more or less unsupportive experiences based on political or cultural views or other social factors. Some participants from marginalized populations might under-report unsupportive experiences to avoid being seen as "whining" or as not earning their success [41, 98]. Alternatively, non-marginalized participants might over-report unsupportive experiences to counter what they see as "woke" popular perceptions [32, 55, 70]. To identify these biases, we include multiple community inclusivity measures and open-ended questions to capture participants' experiences from multiple vantage points. However, these effects likely narrow any differences we might identify between non-marginalized and marginalized populations.

3.6   Positionality Statement

We acknowledge our identities can significantly influence research process and outcomes [6, 48]. Our research team is diverse, comprising three Asian women, three White women (two Jewish), one White nonbinary person, and one White man. The team includes four professional academics who teach security courses, five cybersecurity professionals, four members with government service experience, and two under-

graduate students. All currently reside in the United States. Our overlapping identities as researchers and our personal experiences have led us to observe instances of unwelcoming and unsafe environments in the field of computer security and privacy, as well as instances of harassment in the community.

4 Participants

Table 4 provides a summary of our 342 participants' demographics, divided by gender and race/ethnicity. Most of our participants identified as men (57%) or women (37%). The vast majority of our participants identified as either White or of European descent (63%), Black or of African descent (11%), and/or Hispanic or Latine (13%). Our participants were mostly located in the US (N=279).

The majority of our participants reported having taken at least one programming course in high school (N=234). On average, participants reported 7.5 years of security experience and had job titles including leadership positions, managerial roles, technical positions, and specialized roles related to security analysis and engineering, even holding more senior roles (N=178), such as 'Senior Security Officer" or "CISO". Our participants' had a wide-range of job roles (e.g malware analysis, secure development, and SOC operations).

5 Perceptions of Belonging and Social Experiences (RQ1)

We found few differences in responses between demographics. However, we observed widespread low perceptions of belonging and that women were more likely to experience more severe forms of hate and harassment.

Psychological safety is low for everyone. We observed low psychological safety for all participants: 65.5 on average, which falls into the bottom quartile of scores from a cross-industry benchmark [40]. However, our participants' average belonging uncertainty was lower (indicating less uncertainty) than samples from prior work, which showed higher uncertainty for both non-marginalized and marginalized groups of professionals [122] and students [26]. This suggests our participants overall feel they belong in cybersecurity, but do not feel comfortable speaking up in the community.

No observed differences in perceptions of belonging between genders or races/ethnicities. When comparing psychological safety and belonging among genders the average scores for men (psychological safety 66.6, belonging uncertainty 14.2), women (64.6, 14.3), and genderqueer (60.0, 13.1)were similar. White participants (65.7, 14.2) reported similar scores as participants who identified as Black (64.5, 13.7) or Latine (67.4, 14.6). The similarities across demographic groups can be seen in Figure 2a and Figure 2, which
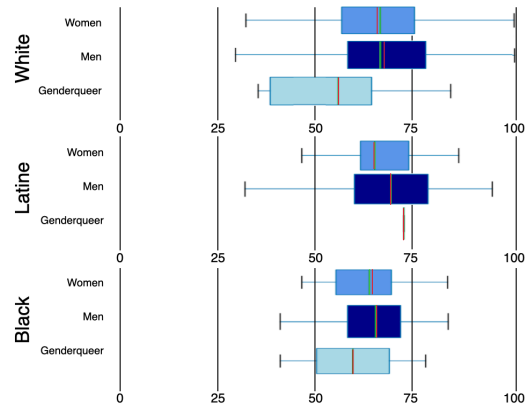
(a) Psychological safety scores across gender and ethnicity
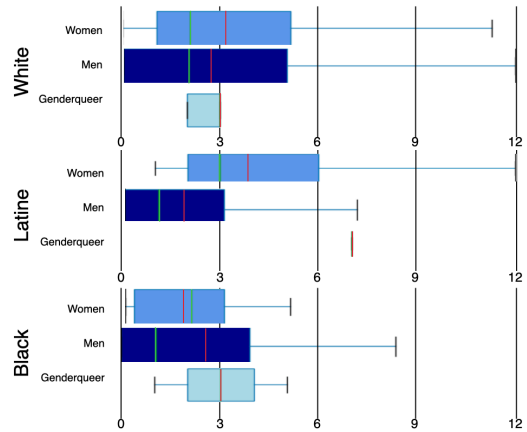
(b) Severe harassment scores across gender and ethnicity

Figure 2: Quantitative results from survey questions about perceptions of belonging and social experiences. The green line indicates the median and the red line indicates the mean for each metric.

plot participants' psychological safety and belonging scores, grouped by gender and race/ethnicity.

The psychological safety regression (Table 5a) found no statistically significant correlation for gender or race/ethnicity. Psychological safety was negatively correlated with participants who reported taking a programming course in high school. Participants who took high school programming scored 5.9 points lower on average while controlling for other factors ($p < 0.001$), indicating participants who began developing cybersecurity skills earlier feel less safe in the community. The final model for belonging uncertainty did not explain a significant variance ($R^2 < 0.02$), so we do not provide it here or discuss it further.

Severe harassment more common for women and those who enter the field earlier. Focusing specifically on severe instances of negative social experiences, namely violence, stereotyping, doxxing, and sexual advances, Figure 2b shows the distribution of severe harassment responses, organized by gender and races/ethnicities. Overall, we note that severe harassment is rare. The average score was 3 out

	Men	**Women**	**Genderqueer**	**White**	**Black**	**Latine**	**Total**
Men	**196 (57%)**	-	-	125 (37%)	22 (6%)	28 (8%)	196 (57%)
Women	-	**128 (37%)**	-	83 (24%)	14 (4%)	17 (5%)	128 (37%)
Genderqueer	-	-	**10 (3%)**	5 (2%)	2 (1%)	1 (0.3%)	10 (3%)
No Answer	-	-	8 (2%)	-	-	-	8 (2%)
White	125 (37%)	83 (24%)	5 (2%)	**215 (63%)**	-	-	215 (63%)
Black	22 (64%)	14 (4%)	2 (1%)	-	**38 (11%)**	-	38 (11%)
Latine	28 (8%)	17 (5%)	1 (0.3%)	-	-	**46 (13%)**	46 (13%)
Avg. Yrs in Sec.	8.3	6.6	5.1	8.1	6.8	8.1	7.5
Heterosexual	169 (49%)	110 (32%)	2 (1%)	186 (54%)	32 (9%)	40 (12%)	291 (85%)
Gay/Lesbian	4 (1%)	3 (1%)	1 (0.3%)	6 (2%)	1 (0.3%)	1 (0.3%)	8 (2%)
Bisexual	8 (2%)	11 (3%)	3 (1%)	12 (4%)	5 (2%)	4 (1%)	24 (7%)
High school	12 (4%)	7 (2%)	0 (0%)	11 (3%)	3 (1%)	2 (1%)	19 (6%)
Some college	15 (4%)	14 (4%)	1 (0.3%)	17 (5%)	5 (2%)	8 (2%)	30 (9%)
Associate degree	12 (4%)	15 (4%)	0 (0%)	21 (6%)	3 (1%)	2 (1%)	27 (8%)
Bachelor's degree	72 (21%)	45 (13%)	5 (2%)	77 (23%)	12 (4%)	22 (6%)	122 (36%)
Master's degree	63 (18%)	39 (11%)	3 (1%)	71 (20%)	14 (4%)	12 (4%)	105 (31%)
Doctorate	14 (4%)	5 (2%)	0 (0%)	17 (5%)	1 (0.3%)	0 (0%)	19 (6%)
Junior role	79 (23%)	43 (13%)	3 (1%)	79 (23%)	13 (4%)	20 (6%)	131 (38%)
Senior role	73 (21%)	55 (16%)	4 (1%)	90 (26%)	18 (5%)	17 (5%)	132 (39%)
Senior leadership	30 (9%)	18 (5%)	0 (0%)	34 (10%)	7 (2%)	8 (2%)	48 (14%)

Table 4: Participant demographics divided by gender identity and race/ethnicity. For each cell we provide the number of participants, as well as the percentage of the total participant pool. Note, we only include the top three most common races/ethnicities and participants could mark multiple races/ethnicities, so those numbers will not sum to 100%. Additionally, two participants self-described their gender identities.

of a possible 12 points and a majority of participants reported "Never" experiencing violence (69.0%), sexual advances (62.0%), or doxxing (60.5%). The exception regarded experiences of stereotype bias, which participants most often reported "Never" experiencing (32.5%), but a non-trivial number reported experiencing it "Rarely" (29.5%), "Occasionally" (25.4%), or Frequently (12.6%).

Women and genderqueer participants reported more frequent occurrences of severe harassment (average frequency scores 3.2 and 3.7 out of 12, respectively) than men (average 2.8). Table 5b shows this correlation was statistically significant ($LE = 1.2, p = 0.015$), indicating women were $1.2\times$ more likely to report more frequent severe harassment than men. We did not observe a similar statistically significant difference for race/ethnicity.

Again, we observed that high school programming experience correlated with an increase in negative outcomes for participants. Participants with high school programming experience were an estimated $1.6\times$ more likely to report severe harassment ($p < 0.001$). We also found participants with more security experience were slightly less likely to report experiences of severe harassment ($LE = 0.9, p = 0.016$).

No observed statistically significant difference in social experiences. We did not observe any statistically significant differences between genders or races/ethnicities on the OSEM scale. Men's (17.9) and women's (17.4) average scores were similar, however genderqueer participants' scores were

slightly higher (21.9), indicating a higher rate of unsupportive experiences. Also, White participants (18.2) reported similar OSEM scores as Black (17.4) or Latine (16.6) participants. The regression over OSEM scores is summarized in in the supplemental materials [101]. The only statistically significant correlation was for participants with a helpful close relationship who were expected to have OSEM scores $0.8\times$ participants' without close relationships ($p < 0.001$). Because this LE is < 1, this indicates a lower score and less negative experiences, as close relationships likely provide important support.

White security experts have lower vulnerability discovery self-efficacy. On average, White participants reported statistically significantly lower vulnerability discovery self-efficacy (32.8) than Black participants (36.6). White participants' scores are estimated to be $0.9\times$ Black participants' scores, holding all other factors equal ($p < 0.001$), as seen in Table 6. We did not see a similar difference between Black and Latine participants. We did not see the same stark differences for gender. On average, men's scores on the vulnerability discovery self-efficacy metric were slightly higher (34.2) than women's (33.0) and genderqueer participants, but gender does not appear in our final regression model.

Security experts with more experience have higher vulnerability discovery self-efficacy. Participants who have left the field or have yet to enter the workforce had lower scores than

Variable	Value	Coef	CI	p
Gender	Man	–	–	–
	Woman	-2.3	[-5.6, 0.9]	0.170
HS Prog	False	–	–	–
	True	**-5.9**	**[-9.4, -2.4]**	**0.001***

– Base case (Coef=0, by definition)
*Significant effect

(a) Psychological safety linear regression.

Variable	Value	LE	CI	p
Gender	Man	–	–	–
	Woman	**1.2**	**[1.6, 3.0]**	**0.015***
Sec Yrs	–	**0.9**	**[0.9, 1.0]**	**0.016***
HS Prog	False	–	–	–
	True	**1.6**	**[1.2, 2.1]**	**<0.001***

– Base case (Log Estimate(LE)=1, by definition)
*Significant effect

(b) Severe harassment Poisson regression.

Table 5: Summary of regression over participant psychological safety (A) and severe harassment (B). R^2 for the psychological safety model was 0.04 and the Pseudo R^2 for the harassment model was 0.06 (corrected Aldrich-Nelson).

those currently working in security ($LE = 0.9$, $p = 0.020$). Participants in more senior ($LE = 1.1$, $p < 0.001$) or C-Suite ($LE = 1.2$, $p < 0.001$) roles reported higher vulnerability discovery self-efficacy than participants in junior roles. Similarly, participants with earlier exposure to programming reported $1.1\times$ higher vulnerability discovery self-efficacy than participants who began programming later ($p = 0.002$). Also, we observed participants who reported having close helpful relationships had statistically significantly higher vulnerability discovery self-efficacy ($LE = 1.1$, $p = 0.012$), echoing the benefits of having a mentor from prior work [38].

6 Subcommunity Participation (RQ2)

Next, we turn to participants' reported subcommunity experiences. Figures 3 and 4 show participants' reported membership in and perception of helpfulness, respectively, for each subcommunity, divided by demographic group. For brevity, we show reported rate of discussion in the supplemental materials [101], as there were no clear differences between groups.

No difference in subcommunity membership or rate of discussion. Participants most often reported discussing security at work (57.3%), having close friends/mentors (56.7%), joining learning communities while in school (48.2%), and participating in online forums or conferences (37.1%). We did not observe a statistically significant difference in subcommunity membership or discussion rates for any demographic groups. Participants who reported joining each subcommunity most often reported discussing security occasionally (54.8% - close friend/mentors, 51.1% - online and conferences, 47.3%

Variable	Value	LE	CI	p
Ethnicity	Black	–	–	–
	Latine	0.9	[0.9,1.0]	0.364
	White	**0.9**	**[0.8,0.9]**	**<0.001***
Close Helpful	False	–	–	–
	True	**1.1**	**[1.0, 1.1]**	**0.011***
HS Prog	False	–	–	–
	True	**1.1**	**[1.0, 1.1]**	**0.002***
Role	Junior	–	–	
	Not in Sec	**0.9**	**[0.7, 1.1]**	**0.020***
	Senior	**1.1**	**[1.1, 1.2]**	**<0.001***
	C-Suite	**1.2**	**[1.1, 1.2]**	**<0.001***

– Base case (Log Estimate (LE)=1, by definition)
*Significant effect

Table 6: Summary of regression over participant vulnerability discovery self-efficacy scores. The Pseudo R^2 (corrected Aldrich-Nelson) for the self-efficacy model was 0.25.

- organizations) or frequently (51.7% - work, 40.1% - school).

Women prioritized identity-focused organizations. Focusing specifically on the types of organizations participants reported being members of, we observed a significant difference between women and men. Women (39.8%) were statistically more likely than men (14.8%) to join "identity-focused" organizations ($\chi^2 = 18.3, p < 0.001$), such as Women in Cybersecurity. However, these rates flipped for general-focus organizations, as men more often joined these groups (59.7%) than women (38.3%), and this difference was statistically significant ($\chi^2 = 6.6, p = 0.010$). We did not observe a similar divide between races/ethnicities for either identity-focused ($\chi^2 = 2.5, p = 0.284$) or general-focus ($\chi^2 = 0.2, p = 0.917$) organizations. We did not observe a similar trend among genderqueer participants as four (of ten) reported membership in identity-focused organizations and four were members of general-focus organizations.

Black participants found community organizations more helpful than White participants. A majority of both Black and White participants who were members of community organizations perceived them as helpful. However, Black participants skewed significantly ($\chi^2 = 5.5, p = 0.019$) more positive regarding community organizations (74.2% extremely helpful, 22.6% somewhat helpful) than White participants (51.3% extremely helpful, 40.3% somewhat helpful).

7 Supportive and Unsupportive Experiences (RQ3)

Finally, we turn to participants' reports of (un)supportive experiences within the cybersecurity community. 307 of the 342 participants responded: 301 described supportive experiences (88.0%) and 291 described unsupportive experiences (85.1%).

We note that a lack of response does not necessarily indicate a lack of supportive or unsupportive experiences, as
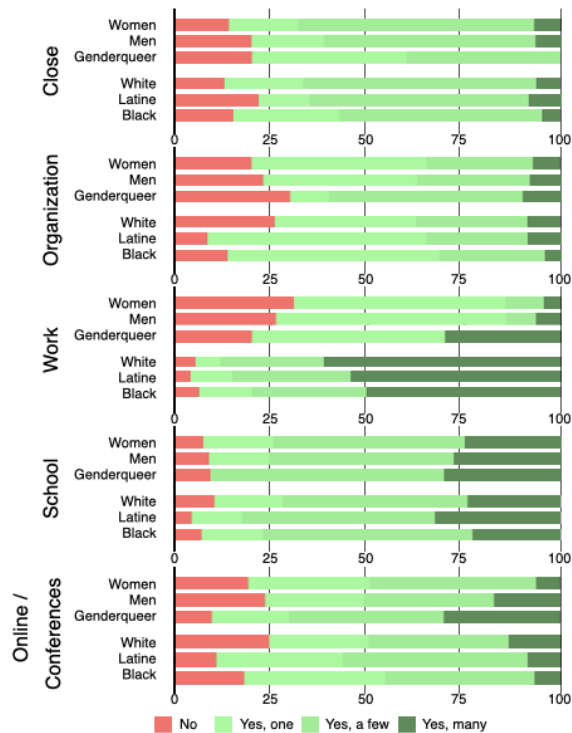
Figure 3: Subcommunity membership grouped by gender and race/ethnicity.

	Men	Women	Genderqueer	Total
White	106 / 101	72 / 73	3 / 3	183 / 179
Black	23 / 21	18 / 18	2 / 2	43 / 41
Latine	25 / 24	17 / 16	1 / 1	43 / 41
Total	172 / 163	115 / 114	7 / 7	301 / 291

Table 7: Participant demographics divided by gender and ethnicity for participants who provided examples of supportive (first number) and unsupportive (second number) community experiences.

responding to these questions was optional. Our analysis focuses on trends observed between men and women, as responses from other genders and races/ethnicities were limited. We did not observe clear differences in the percentages of reported supportive and unsupportive experiences across race/ethnicity, so we do not report those numbers for brevity. However, this should not preclude future work from achieving higher recruitment across these demographics.

Women experienced more unsupportive, negative identity-based incidents. Women (N=14, 12.3%) were significantly ($\chi^2 = 5.18$, $p = 0.023$) more likely to describe encountering an unsupportive environment than men (N=9, 5.5%). For example, one woman shared, "I had a project with a colleague who is not anywhere near as technical as I am, yet he consistently tried to micromanage my technical work, and sometimes told me I was doing things wrong even though he didn't know what he was talking about." When men reported unsupportive environments, these were often due to differing goals
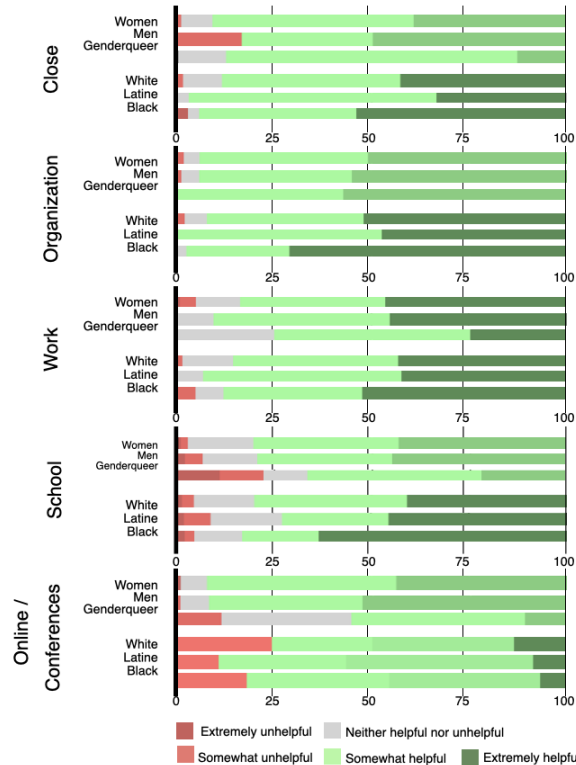
Figure 4: Community helpfulness grouped by gender and race/ethnicity.

or personalities, unlike the devaluation of skills observed with women. For example, one man stated, "[I have been] able to hop projects and or jobs in the past when I felt a workplace was not supportive of the direction I wanted to grow in... Once to avoid policies, and once to avoid a person I did not work with well." Men often reported being able to navigate out of these unsupportive environments relatively easily.

Women (N=16, 14.0%) reported significantly ($\chi^2 = 5.2$, $p = 0.023$) more negative experiences related to their identity than men (N=10, 6.1%). One woman explained, "I was at a career fair and the person at the booth refused to talk to me. They ignored me ... as I patiently waited and proceeded to talk to people who showed up after me... All while refusing to acknowledge my presence. I figure it had something to do with the fact that I was the only female at the booth."

Men more often reported never having an unsupportive experience. Some participants explicitly said they had no supportive (N=10, 3.3%) or unsupportive experiences (N=97, 33.3%). Men were more likely not to have unsupportive experiences (N=59, 36.2%) than women (N=36, 31.6%), though the difference was not statistically significant ($\chi^2 = 1.4$, $p = 0.240$). We did not observe a difference between genders for participants reporting no supportive experiences (6 men and 4 women; $p = 1$ using a Fisher's Exact Test).

Multiple mentions of toxic experiences by both genders.

Bullying, harassment, fear of retaliation, and doxxing were reported by women (N=9, 7.9%) and men (N=14, 8.6%; $\chi^2 = 27$, $p = 0.600$). One man shared he experienced "years of harassment, doxxing, and impersonation," including "fake profiles created in white nationalism and hacking forums." Unwanted attention or sexual advances were mentioned by three women and one man. One participant shared, "A somewhat close friend I had made through a cybersecurity forum had made quite a few uncomfortable sexual remarks which made me question if cybersecurity as a whole was like this or if it was an isolated case."

Men and women reported some common experiences. In addition to the differences discussed above, we also observed some similarities in unsupportive experiences, though these were typically less frequent. These included feeling unwelcome as newcomers (18 men, 11.0%; 6 women, 5.3%; $\chi^2 = 1.3715$, p= 0.242), receiving negative consequences from their own actions (13 men, 8.0%; 10 women, 8.8%; $\chi^2 = 0$, $p = 1$), and difficulty collaborating (22 men, 13.5%; 11 women, 9.6%; $\chi^2 = 0.5$, $p = 0.473$).

There was also general agreement on supportive experiences. Men and women both described receiving career support (41 men, 23.8%; 30 women, 26.1%; $\chi^2 = 0.1$, $p = 0.781$), having their questions answered (26 men, 15.1%; 20 women, 17.4%; $\chi^2 = 0.1$, $p = 0.734$), having positive educational experiences (16 men, 9.3%; 7 women, 6.1%; $\chi^2 = 0.6$, $p = 0.444$), and participating in collaborative problem solving (21 men, 12.2%; 11 women, 9.6%; $\chi^2 = 0.3$, $p = 0.608$).

Women receive support from individuals; men more likely to find groups helpful. Women predominantly cited experiences with specific individuals such as managers, professors, and specific friends or acquaintances (N=47). While the same number of men discussed these individual relationships (N=47), this number is proportionally lower (27.3% of men and 40.9% of women) On the other hand, men discussed supportive experiences with broader groups like online forums or conferences more frequently than women did (34 men, 19.8%; 13 women, 11.3%). One man said, "After sharing a blog post or link to code, someone from the community replied with helpful advice or other areas I could investigate."This difference was statistically significant ($\chi^2 = 6.4$, $p = 0.011$). While we saw a similar divide between men and women in terms of the people involved in unsupportive experiences, i.e., individuals (27 men, 16.6%; 26 women, 22.8%) and groups (27 men, 16.6%; 14 women, 12.3%), this difference was not statistically significant ($\chi^2 = 2.1$, $p = 0.147$).

8 Discussion and Recommendations

Our results provide insights into demographic discrepancies in perceived belonging and community experiences among cybersecurity professionals, alongside overarching trends within the community. We distill common themes from our results and propose actionable recommendations for community and organizational leaders to improve inclusivity and diversity.

8.1 Perceived Belonging and Community Experience Themes

There is a clear gender disparity in community experiences. Across all research questions, the primary divide observed was between genders. Women are more likely to face harassment and unwelcoming experiences related to their identity. Likely due to these unwelcoming experiences, women were less likely to participate in general-interest security organizations, instead opting for identity-based groups, similar to anecdotes presented by Fulton et al. [38].

While we did not have a large enough sample of genderqueer participants to produce generalizable results, these participants' responses suggest they face an unwelcoming community. Across all our survey questions, genderqueer participants reported lower perceptions of belonging and higher experiences of unsupportive environments and identity-based harassment. Future work should focus on this group to better understand the unique challenges they face.

While these experiences mirrored examples described by interview participants in prior work [38], our work demonstrates a clear gender gap and indicates the scale of the problem.

Black participants' responses suggest positive outcomes. Turning to differences between races/ethnicities, we only observed significant differences between Black and White participants. Black participants found community organizations more helpful to their development and career success and had higher vulnerability discovery self-efficacy. This is a positive indicator; however, we remain cautious on this finding as the number of Black participants in our sample was small (N=38) and our findings are only indicative of current cybersecurity professionals' experiences, meaning survivor bias likely plays a role in this result. Further, we stress that while we considered differences in high-level demographic groups, the experiences of members of these groups are not monolithic. Further work is needed to confirm these results with a larger sample and assess the impact of intersectional identities.

Safety perceived as low across participants. Across all demographic groups, we observed low psychological safety scores when compared to results of prior surveys [40]. This lack of perceived safety to engage with others in the community could be internal (e.g., impostor syndrome or perceived lack of knowledge), but our measures of internal belonging and knowledge did not show a similar deficit. This suggests the perceived lack of safety is caused by external forces. which is supported by participants' multiple instances of reported harassment experiences across demographic groups. These results suggest efforts to improve inclusivity and climate in the cybersecurity community would be universally beneficial.

This finding also points to a larger issue of survivor bias in our results. Prior work has shown people with low belonging uncertainty [122] and high self-efficacy [61] are better equipped to overcome negative external forces (like those we observed) because they have a strong internal view of self. Conversely, individuals without the same strong internal perception of self may not join, or remain in, the cybersecurity community due to these negative forces. While our results cannot make claims regarding the people excluded from the community, they point to a potentially high dropout rate and motivate future work investigating this problem.

More work is needed to ensure early cybersecurity education is inclusive and supportive. There have been significant efforts to increase early development programs for cybersecurity-interested students [11, 52, 66, 71, 77, 79, 82]. These programs are important, but our results suggest more work is necessary to investigate and improve their associated communities' inclusivity. Our results indicate cybersecurity professionals who begin skill development early are more likely to face unsupportive environments.

We expected early engagement would lead to stronger perceptions of belonging. While higher vulnerability discovery self-efficacy correlated with high-school programming experience, these early experiences also correlated with lower psychological safety and increased reports of severe harassment, particularly among marginalized groups. This likely contributes to higher dropout rates, emphasizing the need for welcoming and inclusive early education.

While we did not find direct evidence about this, we speculate women are more comfortable participating in gender-specific affinity groups than in general support groups, which may relate to the higher rate of severe harassment women reported, as supported by [38], which described participants avoiding certain groups due to negative experiences.

8.2   Community Leaders Recommendations

Our results indicate a need to improve the culture in cybersecurity to make it more safe and inclusive for everyone, especially women/gender minorities and early career cybersecurity professionals. To this end, we draw on existing best practices from prior work in psychological safety evaluated in other domains [31]. For each best practice, we discuss potential adoption strategies, noting that while these practices were designed for structured workplaces, not all cybersecurity organizations fit this mold (e.g., conference communities, online forums). However, we discuss how the ideas of these practices can be leveraged in less structured environments.

Set the stage. The first step to establish a safer, more inclusive community is for leaders to emphasize safety and inclusivity's value and clearly frame participation in cybersecurity as open to all. This step's goal is to set a shared expectation and vision. For example, #ShareTheMicInCyber promotes the stories and accomplishments of Black cybersecurity professionals, highlighting the impact of their work on the field [99]. Additionally, all major security conferences have established codes of conduct [25, 49, 58, 80, 107, 114] and many have adopted diversity and inclusion statements [23, 50, 53, 115] extolling the importance of a welcoming community, indicating goals for inclusivity, and establishes that hate and harassment that will not be tolerated. While there has been a significant increase in this stage setting recently, it is important that these messages are repeated regularly and within all subcommunities.

Invite participation. While setting the stage is important for creating a shared vision in the community, it is not as meaningful if action is not taken to foster inclusivity. Action is not only the responsibility of leaders, but all members because parts of the cybersecurity community lack clear leaders and structure. Unfortunately, the low psychological safety observed suggests cybersecurity professionals may be less likely to stand up as allies. To counter this issue, community leaders should provide training and support that encourage being an ally and bystander intervention [4], more empathetic responses, and a transition away from a victim-blaming.

Responding productively. Finally, it is paramount that cybersecurity professionals experience actual safety when participating in the community through demonstrated support. The most important practice here is to sanction clear instances of hate and harassment, especially targeting women and genderqueer cybersecurity professionals, as those were seen as most prevalent is our results. This response requires transparent and clear guidelines to avoid silencing expression. Our results indicate cybersecurity professionals experience hate and harassment that crosses a clear boundary, according to most existing policies [23, 25, 49, 50, 53, 58, 80, 107, 114, 115], so it is important these actions are sanctioned publicly to demonstrate the community's commitment to inclusivity. In some subcommunities with less structure, sanctions may be harder to employ, as there is limited central control. These cases demonstrate, again, the need to develop a broad culture of inclusion among community members. For example, it may not be possible for moderators to effectively ban offending users on an anonymous site, as these users can just create new accounts. Allies, instead, might respond with support for the victim and make it clear the offenders' views are not acceptable or representative of the subcommunity.

Our results also suggest destigmatizing mistakes for beginners, especially in early development phases. Beginners may ask easily searchable questions, which can seem frustrating for overworked security educators, but should not be met with disdain [118]. Instead, using resources like FAQs and detailed walkthroughs can help. Questions that persist should be addressed with care, as having someone reliable to ask is crucial for development. Practicing empathy and patience is vital, as most cybersecurity professionals experience some insecurity.

9 Acknowledgements

References

[1] M K Ahuja. Women in the information technology profession: a literature review, synthesis and research agenda. *European Journal of Information Systems*, 11(1):20–34, 2002.

[2] Omer Akgul, Taha Eghtesad, Amit Elazari, Omprakash Gnawali, Jens Grossklags, Michelle L. Mazurek, Aron Laszka, and Daniel Votipka. Bug hunters' perspectives on the challenges and benefits of the bug bounty ecosystem. In *32nd USENIX Security Symposium*, USENIX Sec '23, Los Angeles, CA, August 2022. USENIX Association.

[3] Catherine Ashcraft, Elizabeth Eger, and Michelle Friend. Girls in it: The facts. Technical report, National Center for Women & Information Technology, 2012.

[4] American Psychological Association. Bystander intervention tip sheet. https://www.apa.org/pi/health-equity/bystander-intervention. (Accessed 02-13-2024).

[5] Markus Baer and Michael Frese. Innovation is not enough: climates for initiative and psychological safety, process innovations, and firm performance. *Journal of Organizational Behavior*, 24(1):45–68, 2003.

[6] Shaowen Bardzell and Jeffrey Bardzell. Towards a feminist hci methodology: social science, feminism, and hci. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 675–684, 2011.

[7] Joseph Biden. Executive order on improving the nation's cybersecurity, May 2021. (Accessed 07-21-2021).

[8] Boeing. Boeing technical apprenticeship. https://jobs.boeing.com/btap. (Accessed 01-05-2024).

[9] Amiangshu Bosu and Kazi Zakia Sultana. Diversity and inclusion in open source software (oss) projects: Where do we stand? In *2019 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pages 1–11, 2019.

[10] BugCrowd. Inside the mind of a hacker 2020, 2020. (Accessed 07-21-2020).

[11] Tanner J. Burns, Samuel C. Rios, Thomas K. Jordan, Qijun Gu, and Trevor Underwood. Analysis and exercises for engaging beginners in online CTF competitions for security education. In *2017 USENIX Workshop on Advances in Security Education (ASE 17)*, Vancouver, BC, August 2017. USENIX Association.

[12] A Colin Cameron and Pravin K Trivedi. *Regression analysis of count data*, volume 53. Cambridge university press, 2013.

[13] Sang-Mi Chai and Min-Kyun Kim. A road to retain cybersecurity professionals: An examination of career decisions among cybersecurity scholars. *Journal of the Korea Institute of Information Security & Cryptology*, 22(2):295–316, 2012.

[14] Nina Chamlou. Diversity in cybersecurity. https://www.cyberdegrees.org/resources/diversity-in-cybersecurity/, 2022.

[15] Sapna Cheryan, Benjamin J. Drury, and Marissa Vichayapai. Enduring influence of stereotypical computer science role models on women's academic aspirations. *Psychology of Women Quarterly*, 37(1):72–79, 2013.

[16] Sapna Cheryan, Allison Master, and Andrew N Meltzoff. Cultural stereotypes as gatekeepers: Increasing girls' interest in computer science and engineering by diversifying stereotypes. *Frontiers in psychology*, 6:49, 2015.

[17] Sapna Cheryan, Victoria C Plaut, Paul G Davies, and Claude M Steele. Ambient belonging: how stereotypical cues impact gender participation in computer science. *Journal of personality and social psychology*, 97(6):1045, 2009.

[18] Sapna Cheryan, John Oliver Siy, Marissa Vichayapai, Benjamin J Drury, and Saenam Kim. Do female and male role models who embody stem stereotypes hinder women's anticipated success in stem? *Social Psychological and Personality Science*, 2(6):656–664, 2011.

[19] Debbie Clayton and Teresa Lynch. Ten years of strategies to increase participation of women in computing programs: The central queensland university experience: 1999–2001. *SIGCSE Bull.*, 34(2):89–93, jun 2002.

[20] David A Cole, Elizabeth A Nick, Rachel L Zelkowitz, Kathryn M Roeder, and Tawny Spinelli. Online social support for young people: does it recapitulate in-person social support; can it help? *Computers in human behavior*, 68:456–464, 2017.

[21] Christopher J Collins and Ken G Smith. Knowledge exchange and combination: The role of human resource practices in the performance of high-technology firms. *Academy of management journal*, 49(3):544–560, 2006.

[22] Joel Cooper. The digital divide: The special case of gender. *Journal of computer assisted learning*, 22(5):320–334, 2006.

[23] Cas Cremers and Engin Kirda. Diversity and inclusion. https://www.sigsac.org/ccs/CCS2024/code-of-conduct/diversity-and-inclusion.html. (Accessed 02-02-2024).

[24] Robert Gordon Kent de Grey. *Friends in High-Tech Places: The Development and Validation of the Online Social Support Measure*. PhD thesis, The University of Utah, 2018.

[25] Code of conduct. https://defcon.org/html/links/dc-code-of-conduct.html. (Accessed 02-13-2024).

[26] Anne Deiglmayr, Elsbeth Stern, and Renate Schubert. Beliefs in "brilliance" and belonging uncertainty in male and female stem students. *Frontiers in psychology*, 10:442252, 2019.

[27] James R Detert and Ethan R Burris. Leadership behavior and employee voice: Is the door really open? *Academy of management journal*, 50(4):869–884, 2007.

[28] Michelle Drolet. Diversity in cybersecurity: Barriers and opportunities for women and minorities. https://www.csoonline.com/article/571811/diversity-in-cybersecurity-barriers-and-\opportunities-for-women-and-minorities.html, 2021.

[29] Michael H. Dunn and Laurence D. Merkle. Assessing the impact of a national cybersecurity competition on students' career interests. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, SIGCSE '18, page 62–67, New York, NY, USA, 2018. Association for Computing Machinery.

[30] Amy Edmondson. Psychological safety and learning behavior in work teams. *Administrative science quarterly*, 44(2):350–383, 1999.

[31] Amy C Edmondson. *The fearless organization: Creating psychological safety in the workplace for learning, innovation, and growth*. John Wiley & Sons, 2018.

[32] Rachel Fairbank. Psychologists persevere in edi work despite growing backlash against racial equity efforts. https://www.apa.org/monitor/2024/01/trends-anti-equity-diversity-inclusion-laws, 2024. (Accessed 02-08-2024).

[33] Ronald A Fisher. On the interpretation of χ 2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922.

[34] Cynthia E Foor, Susan E Walden, and Deborah A Trytten. "i wish that i belonged more in this whole engineering group:" achieving individual diversity. *Journal of Engineering Education*, 96(2):103–115, 2007.

[35] National Initiative for Cybersecurity Careers and Studies. Cybersecurity education and training assistant program. https://niccs.cisa.gov/cybersecurity-career-resources/cybersecurity-education-and-training-assistance-program. (Accessed 01-05-2024).

[36] National Initiative for Cybersecurity Education. National initiative for cybersecurity education cybersecurity workforce framework. https://www.cisa.gov/national-initiative-cybersecurity-education-\nice-cybersecurity-workforce-framework. (Accessed 01-05-2024).

[37] Karl Pearson F.R.S. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50(302):157–175, 1900.

[38] Kelsey R. Fulton, Samantha Katcher, Kevin Song, Marshini Chetty, Michelle L. Mazurek, Chloé Messdaghi, and Daniel Votipka. Vulnerability discovery for all: Experiences of marginalization in vulnerability discovery. In *Proceedings of the 44th IEEE Symposium on Security and Privacy*, IEEE S&P '23, 2023.

[39] Kelsey R. Fulton, Daniel Votipka, Desiree Abrokwa, Michelle L. Mazurek, Michael Hicks, and James Parker. Understanding the how and the why: Exploring secure development practices through a course competition. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, CCS '22, New York, NY, USA, 2022. Association for Computing Machinery.

[40] David A Garvin, Amy C Edmondson, and Francesca Gino. Is yours a learning organization? *Harvard business review*, 86(3):109, 2008.

[41] Thomas Gilovich, Dacher Keltner, and Richard E Nisbett. Being a member of a stigmatized group: stereotype threat. *Gilovich, Thomas; Keltner, Dacher; Nisbett, Richard E., Social psychology, New York: WW Norton*, pages 467–468, 2006.

[42] John W. Graham and Steven A. Smith. Gender differences in employment and earnings in science and engineering in the us. *Economics of Education Review*, 24(3):341–354, 2005.

[43] Robert G Kent de Grey, Bert N. Uchino, Brian RW Baucom, Timothy W. Smith, Avery E. Holton, and Edward F. Diener. Enemies and friends in high-tech places: the development and validation of the online social experiences measure. *Digital Health*, 5(1), 2019.

[44] HackerOne. 2019 hacker-powered security report. Technical report, HackerOne, San Francisco, California, December 2019.

[45] HackerOne. The 2020 hacker report. Technical report, HackerOne, San Francisco, California, December 2020.

[46] Andrew F Hayes and Klaus Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89, 2007.

[47] Eoin Hinchy. Voice of the soc analyst. https://www.tines.com/reports/voice-of-the-soc-analyst, 2022.

[48] Andrew Gary Darwin Holmes. Researcher positionality–a consideration of its influence and place in qualitative research–a new researcher guide. *Shanlax International Journal of Education*, 8(4):1–10, 2020.

[49] Code of conduct. https://www.blackhat.com/code-of-conduct.html. (Accessed 02-13-2024).

[50] Diversity and inclusion. https://www.blackhat.com/html/sustainability.html. (Accessed 02-13-2024).

[51] Ohio Cyber Range Insitute. Ohio cyber range institute: Unlocking potential, securing the future. https://www.ohiocyberrangeinstitute.org/. (Accessed 01-06-2024).

[52] SANS Cybersecurity Institute. Girls go cyberstart. (Accessed 05-27-2020).

[53] Diversity and inclusion. https://www.ieee.org/about/diversity-index.html. (Accessed 02-13-2024).

[54] ISC2. Isc2 cybersecurity workforce study 2023. https://media.isc2.org/-/media/Project/ISC2/Main/Media/documents/research/ISC2_Cybersecurity_Workforce_Study_2023.pdf?rev=28b46de71ce24e6ab7705f6e3da8637e, 2023.

[55] Shanto Iyengar, Gaurav Sood, and Yphtach Lelkes. Affect, Not Ideology: A Social Identity Perspective on Polarization. *Public Opinion Quarterly*, 76(3):405–431, 09 2012.

[56] Harjot Kaur, Sabrina Amft, Daniel Votipka, Yasemin Acar, and Sascha Fahl. Where to recruit for security development studies: Comparing six software developer samples. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 4041–4058, 2022.

[57] Robert G Kent de Grey, Bert N Uchino, Brian RW Baucom, Timothy W Smith, Avery E Holton, and Edward F Diener. Enemies and friends in high-tech places: the development and validation of the online social experiences measure. *Digital Health*, 5:2055207619878351, 2019.

[58] Engin Kirda and David Lie. Code of conduct. https://www.sigsac.org/ccs/CCS2024/code-of-conduct/code-of-conduct.html. (Accessed 02-02-2024).

[59] Solomon Klappholz. A third of cyber security pros report crumbling work-life balance. https://www.itpro.com/security/a-third-of-cyber-security-pros-report-crumbling-\work-life-balance, 2023.

[60] Andrew J Ko. Attitudes and self-efficacy in young adults' computing autobiographies. In *2009 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pages 67–74. IEEE, 2009.

[61] Robert W. Lent, Steven D. Brown, and Gail Hackett. Toward a unifying social cognitive theory of career and academic interest, choice, and performance. *Journal of Vocational Behavior*, 45(1):79 – 122, 1994.

[62] Jian Liang, Crystal I. C. Farh, and Jiing-Lih Farh. Psychological antecedents of promotive and prohibitive voice: A two-wave examination. *Academy of Management Journal*, 55(1):71–92, 2012.

[63] Thomas Maillart, Mingyi Zhao, Jens Grossklags, and John Chuang. Given enough eyeballs, all bugs are shallow? revisiting eric raymond with bug bounty programs. In *Proceedings of the 15th Workshop on the Economics of Information Security*, WEIS '16, 2016.

[64] Jane Margolis, Rachel Estrella, Joanna Goode, Jennifer Jellison Holme, and Kim Nao. *Stuck in the shallow end: Education, race, and computing*. MIT press, 2017.

[65] Jane Margolis and Allan Fisher. *Unlocking the clubhouse: Women in computing*. MIT press, Cambridge, MA, 2002.

[66] Luke McCormack. Us cyber challenge. https://www.uscyberchallenge.org/, September 2022. (Accessed 09-17-2022).

[67] Robert K McKinley, Terjinder Manku-Scott, Adrian M Hastings, David P French, and Richard Baker. Reliability and validity of a new measure of patient satisfaction with out of hours primary medical care in the united kingdom: development of a patient questionnaire. *Bmj*, 314(7075):193, 1997.

[68] Adam W Meade and S Bartholomew Craig. Identifying careless responses in survey data. *Psychological methods*, 17(3):437, 2012.

[69] Uta Menges, Jonas Hielscher, Annalina Buckmann, Annette Kluge, M Angela Sasse, and Imogen Verret. Why it security needs therapy. In *European Symposium on Research in Computer Security*, pages 335–356. Springer, 2021.

[70] Rachel Minkin. Personal experiences with online harassment. https://www.pewresearch.org/social-trends/2023/05/17/diversity-equity-and-inclusion-in-the-workplace/, 2023. (Accessed 02-08-2024).

[71] Jelena Mirkovic and Peter A. H. Peterson. Class capture-the-flag exercises. In *2014 USENIX Summit on Gaming, Games, and Gamification in Security Education (3GSE 14)*, San Diego, CA, August 2014. USENIX Association.

[72] Phoenix Moorman and Elizabeth Johnson. Still a stranger here: Attitudes among secondary school students towards computer science. In *Proceedings of the 8th Annual Conference on Innovation and Technology in Computer Science Education*, ITiCSE '03, page 193–197, New York, NY, USA, 2003. Association for Computing Machinery.

[73] Laurie A. Morgan. Glass-ceiling effect or cohort effect? a longitudinal study of the gender earnings gap for engineers, 1982 to 1989. *American Sociological Review*, 63(4):479–493, 1998.

[74] Dawn Nafus. 'patches don't have gender': What is not open in open source software. *New Media & Society*, 14(4):669–683, 2012.

[75] Elizabeth A Nick, David A Cole, Sun-Joo Cho, Darcy K Smith, T Grace Carter, and Rachel L Zelkowitz. The online social support scale: measure development and validation. *Psychological assessment*, 30(9):1127, 2018.

[76] SH Nielsen, LA von Hellens, and Sharon Wong. The male it domain: You've got to be in it to win it. In *Proceedings of the 12th Australasian Conference on Information Systems (ACIS 2001)*, pages 1–12, 2001.

[77] NIST. Cybersecurity competitions | nist. https://www.nist.gov/itl/applied-cybersecurity/nice/community/community-coordinating-council/cybersecurity-skills-0, 2023. (Accessed 02-08-2024).

[78] Samantha Nix and Lara Perez-Felkner. Difficulty orientations, gender, and race/ethnicity: An intersectional analysis of pathways to stem degrees. *Social Sciences*, 8(2), 2019.

[79] Plaid Parliament of Pwning. picoctf. https://picoctf.com/. (Accessed 05-27-2020).

[80] IEEE Symposium on Security and Privacy. Code of conduct. https://www.ndss-symposium.org/ndss-code-of-conduct/. (Accessed 02-13-2024).

[81] Zhen Xin Ong, Liz Dowthwaite, Elvira Perez Vallejos, Mat Rawsthorne, and Yunfei Long. Measuring online wellbeing: a scoping review of subjective wellbeing measures. *Frontiers in psychology*, 12:616637, 2021.

[82] PACTF. Pactf. https://2019.pactf.com/. (Accessed 05-27-2020).

[83] Toni C Plato. Women c-suite executives in cybersecurity: Transformational experiences and gender barriers on their leadership journeys, 2021.

[84] Portia Pusey, Mark Gondree, and Zachary Peterson. The outcomes of cybersecurity competitions and implications for underrepresented populations. *IEEE Security & Privacy*, 14(6):90–95, 2016.

[85] Laura Quintana. Hack your way to a new career in cybersecurity: Cisco networking academy offers new ethical hacker course. https://blogs.cisco.com/learning/hack-your-way-to-a-new-career-in-cybersecurity, October 2023. (Accessed 01-05-2024).

[86] Jessica Rafaeil. How three stem leaders forged a path for others through fearlessness. *US Black Engineer and Information Technology*, 44(1):22–25, 2020.

[87] Adrian E Raftery. Bayesian model selection in social research. *Sociological methodology*, pages 111–163, 1995.

[88] Laurie A Rudman. Self-promotion as a risk factor for women: the costs and benefits of counterstereotypical impression management. *Journal of personality and social psychology*, 74(3):629, 1998.

[89] David L Sackett. Bias in analytic research. In *The case-control study consensus and controversy*, pages 51–63. Elsevier, 1979.

[90] Robert J Sampson. Local friendship ties and community attachment in mass society: A multilevel systemic model. *American sociological review*, pages 766–779, 1988.

[91] Koen Schoenmakers, Daniel Greene, Sarah Stutterheim, Herbert Lin, and Megan J Palmer. The security mindset: characteristics, development, and consequences. *Journal of Cybersecurity*, 9(1):tyad010, 2023.

[92] Elaine Seymour and Nancy M. Hewitt. *Talking About Leaving: Why Undergraduates Leave the Sciences*. Westview Press, 2000.

[93] Jenessa R. Shapiro and Amy M. Williams. The role of stereotype threats in undermining girls' and women's performance and interest in stem fields. *Journal on Sex Roles*, 66(3):175–183, 2012.

[94] Enno Siemsen, Aleda V Roth, Sridhar Balasubramanian, and Gopesh Anand. The influence of psychological safety and confidence in knowledge on employee knowledge sharing. *Manufacturing & service operations management*, 11(3):429–447, 2009.

[95] Daryl G Smith. *Diversity's Promise for Higher Education: Making It Work*. JHU Press, 2020.

[96] Elliott Sober. Instrumentalism, parsimony, and the akaike framework. *Philosophy of Science*, 69(S3):S112–S123, 2002.

[97] David M. Sparks, Steve Daniel Przymus, Allison Silveus, Yohanis De La Fuente, and Cassandra Cartmill. Navigating the intersectionality of race/ethnicity, culture, and gender identity as an aspiring latina stem student. *Journal of Latinos and Education*, 22(4):1355–1371, 2023.

[98] Claude M. Steele, Steven J. Spencer, and Joshua Aronson. Contending with group image: The psychology of stereotype and social identity threat. volume 34 of *Advances in Experimental Social Psychology*, pages 379–440. Academic Press, 2002.

[99] Camille Stewart, Lauren Zabierek, and Katelyn Ringrose. #sharethemicincyber. https://www.sharethemicincyber.com/.

[100] Anselm Strauss and Juliet Corbin. *Basics of qualitative research*, volume 15. Newbury Park, CA: Sage, 1990.

[101] Supplemental materials. https://osf.io/k3m8g.

[102] Synack. Synack cybersecurity diversity and inclusion report. https://www.synack.com/diversity-report/, 2020.

[103] Mohammad Tahaei, Ruba Abu-Salma, and Awais Rashid. Stuck in the permissions with you: Developer & end-user perspectives on app permissions & their privacy ramifications. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–24, 2023.

[104] Mohammad Tahaei and Kami Vaniea. Recruiting participants with programming skills: A comparison of four crowdsourcing platforms and a cs student mailing list. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2022.

[105] Josh Terrell, Andrew Kofink, Justin Middleton, Clarissa Rainear, Emerson Murphy-Hill, Chris Parnin, and Jon Stallings. Gender differences and bias in open source: Pull request acceptance of women versus men. *PeerJ Computer Science*, 3:e111, 2017.

[106] Steven Terrell and Kembley Lingelbach. A study of female cybersecurity professionals. *Issues in Information Systems*, 24(3), 2023.

[107] Code of conduct. https://www.ndss-symposium.org/ndss-code-of-conduct/. (Accessed 02-13-2024).

[108] Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley, Deepak Kumar, Damon McCoy, Sarah Meiklejohn, Thomas Ristenpart, and Gianluca Stringhini. Sok: Hate, harassment, and the changing landscape of online abuse. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 247–267, 2021.

[109] Eileen M Trauth. Mapping information-sector work to the work force. *Communications of the ACM*, 44(7):74–75, 2001.

[110] Thomas Trevethan and Grace Williams. The time is now to secure the future. https://www.paloaltonetworks.com/blog/2023/10/secure-the-future/, October 2023. (Accessed 01-06-2024).

[111] Chih-Hsiung Tu. The measurement of social presence in an online learning environment. In *International Journal on E-learning*, volume 1, pages 34–45. Association for the Advancement of Computing in Education (AACE), 2002.

[112] Sherry Turkle. *The Second Self: Computers and the Human Spirit*. Mit Press, 1984.

[113] Visa University. Visa payments cybersecurity certification. https://www.visauniversity.com/en/certificate-programs/visa-certification/visa-payments-cybersecurity-certification. (Accessed 01-06-2024).

[114] Code of conduct. https://www.usenix.org/conferences/coc. (Accessed 02-13-2024).

[115] Diversity and inclusion. https://www.usenix.org/conferences/diversity-and-inclusion. (Accessed 02-13-2024).

[116] Bogdan Vasilescu, Daryl Posnett, Baishakhi Ray, Mark G.J. van den Brand, Alexander Serebrenik, Premkumar Devanbu, and Vladimir Filkov. Gender and tenure diversity in github teams. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 3789–3798, New York, NY, USA, 2015. Association for Computing Machinery.

[117] Emily A. Vogels. Personal experiences with online harassment. https://www.pewresearch.org/internet/2021/01/13/personal-experiences-with-online-harassment/, 2021. (Accessed 01-06-2024).

[118] D. Votipka, E. Zhang, and M. Mazurek. Hacked: A pedagogical analysis of online vulnerability discovery exercises. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 1589–1606, Los Alamitos, CA, USA, may 2021. IEEE Computer Society.

[119] Daniel Votipka, Desiree Abrokwa, and Michelle L. Mazurek. Building and validating a scale for secure software development self-efficacy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–20, New York, NY, USA, 2020. Association for Computing Machinery.

[120] Daniel Votipka, Hongyi Hu, Bryan Eastes, and Michelle L. Mazurek. Toward a field study on the impact of hacking competitions on secure development. In *Proceedings of the 4th Workshop on Security Information Workers*, WSIW '18, Baltimore, MD, 2018. USENIX Association.

[121] Daniel Votipka, Rock Stevens, Elissa M Redmiles, Jeremy Hu, and Michelle L Mazurek. Hackers vs. testers: A comparison of software vulnerability discovery processes. In *Proceedings of the 39th IEEE Symposium on Security and Privacy*, IEEE S&P '18, 2018.

[122] Gregory M Walton and Geoffrey L Cohen. A question of belonging: race, social fit, and achievement. *Journal of personality and social psychology*, 92(1):82, 2007.

[123] Miranda Wei, Pardis Emami-Naeini, Franziska Roesner, and Tadayoshi Kohno. Skilled or gullible? gender stereotypes related to computer security and privacy. In *IEEE Symposium on Security and Privacy*, 2023.

[124] Elizabeth Williams. Actualizing gender and racial diversity inclusion in computing fields. *Issues in Information Systems*, 24(4), 2023.

[125] Women in security and privacy. https://www.wisporg.com/, 2024.

[126] Danielle M. Young, Laurie A. Rudman, Helen M. Buettner, and Meghan C. McLean. The influence of female role models on women's implicit science cognitions. *Psychology of Women Quarterly*, 37(3):283–292, 2013.

A Survey

Helpfulness Scale. Extremely helpful, Somewhat helpful, Neither helpful nor unhelpful, Somewhat unhelpful, Extremely unhelpful

Agreement Scale. Strongly disagree, Disagree, Somewhat disagree, Neither agree nor disagree, Somewhat agree, Agree, Strongly agree

A.1 General Belonging

This section asks about your experiences in the cybersecurity community and any support in these settings you have been provided toward your security education and career.

For the following statements, please indicate the extent to which they reflect your experience in the cybersecurity community. *1 (Very inaccurate), 2, 3, 4 (Neither accurate nor inaccurate), 5, 6, 7 (Very accurate)*

1. In cybersecurity spaces, it is easy to speak up about what is on your mind.

2. If you make a mistake in the cybersecurity community, it is often held against you.

3. People in cybersecurity are usually comfortable talking about problems and disagreements.

4. People in cybersecurity are eager to share information about what doesn't work as well as share information about what does work.

5. Keeping your cards close to your chest is the best way to get ahead in the cybersecurity community.

Please indicate your level of agreement with the following statements: *Agreement Scale*

1. Sometimes I feel that I belong in cybersecurity, and sometimes I feel that I don't belong.

2. When something bad happens, I feel that maybe I don't belong in cybersecurity.

3. When something good happens, I feel that I really belong in cybersecurity.

4. People from different backgrounds have equal opportunities to participate in the cybersecurity community.

For the following items, think about your interactions with your professors/peers/collegues in the computer security community. To respond, indicate to what extent you felt this way. *Very Slightly or Not at all, A Little, Moderately, Quite a Bit, Extremely*

1. Interactions with someone in the field prevented me from working on my goals or other important things.

2. Someone in the cybersecurity community has encouraged me when I felt like quitting.

3. I have felt supported by someone in the cybersecurity community who agreed with my point of view.

4. I have been unable to fall asleep while thinking about a negative interaction I had with someone in the cybersecurity community.

5. There are people in the cybersecurity community please ignore the first part of this statement and mark "Extremely".

6. Someone in the cybersecurity community has cheered me up when I was feeling down.

7. Someone in the cybersecurity community has made me feel embarrassed or foolish.

8. There is someone in the cybersecurity community I can turn to for advice about handling problems.

9. There is someone in the cybersecurity community I could turn to for advice about making career plans or about changing my job.

Have you ever experienced any of the following behaviors directed at you in the context of the cybersecurity community? *Never, Rarely, Occasionally, Frequently*

1. Lack of response or rejection of contributions or questions.

2. Conflict or interpersonal tension between you and another community member.

3. Written or spoken language that made you feel unwelcome (e.g. profanity, racist jokes, sexual imagery, hostility, rudeness, name calling).

4. Stereotyping based on perceived demographic characteristics.

5. Threats of violence, stalking.

6. Unsolicited sexual advance or comments.

7. Impersonation or malicious publication of personal information (doxxing).

A.2 Close Relations

This section asks about your experience with family, friends and other close mentors, and any support they have provided toward your security education and career.

1. Do you have an family, friends or other close mentors you go to for help when you're trying to learn difficult security concepts? *Yes, one; Yes, a few; Yes, many; No*

2. How often do you discuss technical topics related to your security education and career with your family/friends/mentors? *Never, Rarely, Occasionally, Frequently*

3. How helpful do you find the security-related guidance that your family/friends/mentors give you? *Helpfulness scale*

A.3   Workplace

This section asks about your experiences in your workplace and any support in these settings you have been provided toward your security education and career.

1. Are you a part of any workplaces where security concepts are discussed? *Yes, but security concepts are rarely discussed; Yes, security concepts are sometimes discussed; Yes, security concepts are often discussed; No, but I was previously employed in security; No*

2. What is the primary focus of the company you work for? *Non-technical - critical infrastructure (hospitals, power, etc.), Non-technical - non-critical infrastructure, Security - defense (intrusion detection/response, system defense/hardening), Security - offense (penetration testing, vulnerability analysis), Other technical - software development, Other technical - network/system administration, Other technical - hardware development, Other*

3. How often do you discuss technical topics related to your security education and career in your workplace? *Never, Rarely, Occasionally, Frequently*

4. How helpful do you find the security-related guidance given to you by colleagues in your workplace? *Helpfulness scale*

A.4   Organizations

This section asks about your experience with any organizations you participate in and any support you have received toward your security education or career. We consider an organization as any group outside your work/classes where people meet regularly to discuss technical topics of interest (e.g., local ACM chapter, Women in Security and Privacy). Please answer the following questions only considering security-related organizations.

1. Are you a part of any organizations where security concepts are discussed? *Yes, I participate in one organization where security concepts are discussed; Yes, I participate in a few organizations where security concepts are discussed; Yes, I participate in many organizations where security concepts are discussed; No*

2. What kinds of organizations are you a part of? *"Identity-based (e.g. Women in Security, LGBTQ+ in Security, Blacks in Cyber)", "Topic-based (e.g. malware analysis working group)", "General security group", "Other"*

3. How often do you discuss technical topics related to your security education and career with people in these organizations? *Never, Rarely, Occasionally, Frequently*

4. How helpful do you find the security-related guidance given to you by people in organizations you are a part of? *Helpfulness scale*

A.5   School

This section asks about your experiences in your school and any support in these settings you have been provided toward your security education and career. This section only pertains to academic situations, e.g., classes, professors, peers, organizations.

1. Have you taken any classes where security concepts are discussed? *Yes, one; Yes, a few; Yes, many; No*

2. Are you a part of any school organizations where security concepts are discussed? *Yes, I participate in one organization where security concepts are discussed; Yes, I participate in a few organization where security concepts are discussed; Yes, I participate in many organization where security concepts are discussed; No*

3. How often do you discuss technical topics related to your security education in your school? *Never, Rarely, Occasionally, Frequently*

4. How helpful do you find the security-related guidance given to you by professors and peers? *Helpfulness scale*

A.6   Broader Security Community

This section asks about your experience with the broader security community, including conferences/workshops or online when asking questions about or discussing computer security topics. We define online community discussions as any discussion about security concepts in a public online forum (e.g., StackOverflow, Reddit, Twitter, public Slack or Discord).

1. Have you participated in the broader security community (conferences/workshops or online security communities)?

   '*Yes, I participate in one public conference or online community; Yes, I participate in a few public conferences or online communities; Yes, I participate in many public conferences or online communities; No*

2. Please indicate any security conferences or workshops you have participated in.

3. Please indicate any forums or social media platforms you use for interacting with the public online security community.

4. How often do you discuss technical topics related to your security education and career with people in the broader security community? *Never, Rarely, Occasionally, Frequently*

5. How helpful do you find the security-related guidance given to you by people in the broader security community? *Helpfulness scale*

A.7 General

We have asked several questions pertaining to your experiences with various communities surrounding you—close contacts, school/workplaces, organizations, and the broader security community. For the next section, we'll ask you to consider all the experiences you've had with others in the security community. For both questions, we ask that you do not name specific individuals.

1. Please describe one particularly good experience you had with a community you are part of (and mention which community—close contacts, school/workplaces, organizations, or the broader security community). This could be any experience where you felt the other individual was supportive and helped your development or career in a tangible or intangible way.

2. Please describe one particularly bad experience you had with a community you are part of (and mention which community—close contacts, school/workplaces, organizations, or the broader security community). This could be any experience where you felt the other individual was not supportive and the interaction was harmful to your development or career in a tangible or intangible way.

Now we will ask some questions pertaining to your background and experience in computer security.

1. Do you work in a role where you are asked to perform security tasks? *Yes, this is the primary focus of my job; Yes, this is a part of my job, but not the primary focus; I previously worked in a role where security was my primary focus; I previously worked in a role where security was part of the job; No*

2. Please indicate the approximate number of years you have worked in security.

3. What is/was your position title?

4. Did you choose to leave? *Yes, No*

5. If you feel comfortable sharing, what were your reasons for leaving?

Please indicate your level of agreement with the following statements: *Agreement Scale*

1. I am interested in continuing my security education.

2. I am interested in pursuing or continuing to pursue a career in security.

3. I am well-prepared for a career in security.

Please indicate:

1. Have you participated in any of the following types of security education? (Select all that apply) *Capture-the-flag, wargames, or other online security competitions (e.g., picoCTF, crackmes, iCTF), Penetration testing lab (e.g., Hack the box) or cyber range exercise, Professional certification course (e.g., GIAC Security Essentials, Certified Ethical Hacker), Conference workshop (e.g., Defcon Village workshops), MOOC security course (e.g., Coursera Cybersecurity Specialization), Academic course, Other, I have not participated in any security education*

2. Where did you typically rank when participating in CTFs or other online security competitions? *Top 25% of participants, 25-50% of participants, 50-75% of participants, Bottom 25% of participants*

Please indicate how confident you are in the following statements: *Not confident at all, Slightly confident, Somewhat confident, Moderately confident, Absolutely confident*

1. I can perform a threat risk analysis (e.g., likelihood of vulnerability, impact of exploitation)

2. I can identify potential security threats to the system

3. I can identify the common attack techniques used by attackers

4. I can identify potential attack vectors in the environment the system interacts with (e.g., hardware, libraries)

5. I can identify common vulnerabilities of a programming language

6. I can identify the common please ignore this question and select "Absolutely confident"

7. I can design software to quarantine an attacker if a vulnerability is exploited

8. I can mimic potential threats to the system

9. I can evaluate security controls on the system's interfaces/interactions with other software systems

10. I can evaluate security controls on the system's interfaces/interactions with hardware systems

Cybersecurity development experiences

1. When was the earliest time you remember first being interested in computer security? Please indicate your approximate age: *(Number)*

2. When was the earliest time you had someone (e.g., friend, family member, colleague) in your life who you could go to for security education support? Please indicate your approximate age: *(Number)*

3. Aside from direct educational support, do you have anyone (e.g., friend, family member, colleague) who support your educational pursuits in security (e.g., encouragement, monetary support)? *Yes, I have one person who has provided non-educational support; Yes, I have a few people who have provided non-educational support; Yes, I have many people who have provided non-educational support; No*

4. When was the earliest time you had someone (e.g., friend, family member, colleague) in your life who supported your pursuit of a security education aside from direct teaching? Please indicate your approximate age: *(Number)*

A.8   Demographics

1. In which country do you currently reside?

2. How old are you? *18-19, 20-24, 25-29, 30-35, 35-39, 40-44 , 45-49, 50-54, 55+, Prefer not to answer*

3. What is your ethnicity? *White or of European descent, South Asian, Hispanic or Latino/a/x, Middle Eastern , East Asian, Black or of African descent, Southeast Asian, Indigenous (such as Native American, Pacific Islander, or Indigenous Australian), Prefer to self-describe, Prefer not to answer*

4. What is your gender? *Woman, Man, Transgender Woman / Trans Feminine, Transgender Man / Trans Masculine,* *Non-Binary / Genderqueer / Gender Fluid, Two Spirit, Prefer to state, Prefer not to answer*

5. What is your sexual orientation? Do you identify as: *Bisexual, Gay/Lesbian, Heterosexual/Straight, Don't know, Prefer to self-describe, Prefer not to say*

6. What is the highest degree or level of school you have completed? *High school, Some college or currently enrolled, Associate's degree, Bachelor's degree, Master's/ Professional degree, Doctorate degree, Prefer not to say*

7. Did you take any programming classes or training in high school? *Yes, one; Yes, a few; Yes, many; No*

8. Which range matches most closely your total, pre-tax household income over the last fiscal year? *< $29,999 , $30,000 - $49,999, $50,000 - $74,999, $75,000 - $99,999, $100,000 - $124,999 , $125,000 - $149,999, $150,000 - $174,999 , $175,000 - $199,999 , > $200,000, Prefer not to answer*

9. Which range matches most closely your total, pre-tax household income when growing up? (before 18 years old)? *Same as Question 8*

A.9   Final

1. If you like, we may contact you for one of the following reasons. Please indicate what you like to be contacted for (you may select multiple): *Follow-up interview (i.e., questions related to this study); Future research (i.e., questions related to other computer security topics); Raffle for one of 25 $50 Amazon gift cards; None of the above*

2. Please provide your email address so we can contact you for the reasons selected previously. If you chose to only be contacted for the raffle, your email address will be deleted after the raffle has been completed. Your email will not be used for any purpose beyond those you indicated in the previous question.

Evaluating the Usability of Differential Privacy Tools with Data Practitioners

Ivoline C. Ngong
University of Vermont

Brad Stenger
University of Vermont

Joseph P. Near
University of Vermont

Yuanyuan Feng
University of Vermont

Abstract

Differential privacy (DP) has become the gold standard in privacy-preserving data analytics, but implementing it in real-world datasets and systems remains challenging. Recently developed DP tools aim to make DP implementation easier, but limited research has investigated these DP tools' usability. Through a usability study with 24 US data practitioners with varying prior DP knowledge, we evaluated the usability of four open-source Python-based DP tools: DiffPrivLib, Tumult Analytics, PipelineDP, and OpenDP. Our study results suggest that these DP tools moderately support data practitioners' DP understanding and implementation; that Application Programming Interface (API) design and documentation are vital for successful DP implementation and user satisfaction. We provide evidence-based recommendations to improve DP tools' usability to broaden DP adoption.

1 Introduction

Advances in big data analytics have propelled the collection and processing of massive amounts of data, including sensitive data such as medical records, financial information, and other personally identifiable information. The analysis of this sensitive data may result in the accidental leakage of individuals' data [40, 56], even when anonymization techniques are used [15, 16, 34, 60]. Differential privacy (DP) can mitigate these risks [24, 25] by guaranteeing the results of statistical analyses will not reveal too much personal information about any individual. By adding carefully calibrated noise to data, DP protects sensitive data while still revealing high-level statistical insights. Due to its tremendous potential to revolutionize privacy-preserving data analysis, DP attracts considerable research [25]. Leading government organizations and technology companies, including the U.S. Census Bureau [64], Google [29], Apple [6], and Microsoft [44] have also adopted DP to protect individuals' data privacy.

However, current DP adoption is limited outside of large organizations and companies [20], primarily because implementing DP from scratch is complex and error-prone [37]. DP implementations must carefully account for the privacy budget, generate appropriate random noise, and require systems to be safe against known side-channel vulnerabilities. Additionally, scaling these systems to real-world datasets often requires significant software engineering effort.

To address these challenges, various tools, frameworks, and libraries [21, 22, 27, 30, 36, 41, 52–54, 57, 58, 65, 68, 69] (collectively called "DP tools" hereafter) have been developed to make DP implementation accessible to **data practitioners** — defined in this paper as professionals who have data analysis and programming skills but may not be familiar with DP. These DP tools intend to help data practitioners implement DP solutions without privacy failures. Currently, no research has systematically evaluated the usability of these DP tools; therefore, it remains unclear if they truly enable data practitioners to effectively implement DP solutions. If not, usability may be the bottleneck for wider DP adoption.

In this study, we have assessed four open-source Python-based DP tools through a mixed-methods usability study with 24 US data practitioners, evaluating four widely-used usability criteria–learnability, efficiency, error prevention, and user satisfaction [51] – to investigate three research questions:

- How effectively can DP tools help data practitioners understand DP concepts? (RQ1: DP Understanding)

- How effectively can DP tools help data practitioners implement DP solutions? (RQ2: DP Implementation)

- How satisfied are data practitioners with DP tools for their DP implementation? (RQ3: User Satisfaction)

We conducted the first comprehensive cross-tool usability study of four Python-based DP tools with data practitioners. The focus on data practitioners—the potential adopters of DP—enriches the currently end user-centered DP user research. Our contribution lies in the identification of these DP tools' usability issues and in our recommendations to improve DP tools' usability to facilitate broader DP adoption.

2 Related Work

DP and Implementation Challenges. Differential privacy (DP) [24, 25] is a formal privacy definition designed to allow statistical analysis while protecting information about individuals. Differentially private analyses, often called *mechanisms*, typically add random noise to analysis results in order to achieve privacy. Formally, two datasets $D, D' \in \mathcal{D}$ are called *neighboring datasets* if they differ in one person's data, and a mechanism $\mathcal{M}$ satisfies (ε, δ)-DP if for all neighboring datasets D and D' and sets of outcomes S:

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\varepsilon} \Pr[\mathcal{M}(D') \in S] + \delta$$

The ε parameter is the *privacy parameter* or *privacy budget*; a smaller ε results in stronger privacy, while a larger ε results in weaker privacy. Noise drawn from the Laplace or Gaussian distributions can be used to achieve differential privacy.

Existing DP Tools. Implementing DP mechanisms is challenging. Data practitioners must determine the amount of noise to add, limit the total privacy budget, and ensure the system is free of common DP bugs [17, 32, 35, 45]). Numerous tools and libraries have attempted to make implementing DP easier for data practitioners [21, 22, 27, 30, 36, 41, 52–54, 57, 58, 65, 68, 69], often by handling the tricky parts of DP automatically. For example, tools may calculate sensitivity automatically [41, 58, 69] and ensure the privacy budget is not violated [22, 27, 36, 41, 52–54, 65, 68]. They may provide vetted implementations of basic DP mechanisms like the Laplace mechanism [21, 30, 52, 53, 68], and some also support machine learning applications [21, 53, 68]. Notably, DPCreator [22] and Private data Sharing Interface (PSI) [27] provide graphical interfaces designed for non-experts; the other tools require data science knowledge but reduce the need for DP expertise.

User Research around DP Understanding. Existing user research around DP understanding mostly focuses on end users, whose data would be in a differentially private dataset. Bullek et al. [14] examined if animated spinners can effectively communicate DP privacy guarantees to end users. Their participants preferred spinners with higher privacy levels but did not fully trust the spinners. Cummings et al. [19] studied how various DP explanations impact end-user perceptions. They found DP explanations raised participants' expectations of privacy, but did not increase their willingness to share data. Other studies explored how to better communicate DP concepts to end users. Xiong et al. [66] assessed the use of scenarios to communicate the privacy guarantees of three different DP models with participants from the USA and India. Kühtreiber et al. [38] replicated this study with participants from Germany. Both studies indicated that end users lack understanding of DP and highlighted a need for more effective DP communication. German participants were more willing to share data compared to those in the USA and India. Ashena et al. [7] also found interactive visual tools helped communicate the trade-off between accuracy and privacy loss. These studies suggest that end users have difficulty understanding

DP and reservations towards DP's privacy protection.

Currently, limited user research has examined the perspective of data practitioners, who particularly need adequate DP understanding to correctly implement it. One notable study by Nanayakkara et al. [49] tested an interactive interface called Visualizing Privacy (ViP) with data practitioners without DP background, and found visualizing relationships between ε, accuracy, and disclosure risk helped them judge DP noise. Our study extends this line of research to investigate if DP tools could assist data practitioners' DP understanding.

Usability around DP Implementation. Garrido et al. [28] interviewed 24 privacy practitioners and identified both organizational and technical challenges for DP implementation in the industry. Their findings suggested that API-based DP tools could streamline data access integration and DP implementation across the enterprise.

A few studies have evaluated the usability of specific DP tools. Murtagh et al. [46] studied the usability of the web-based DP tool Privacy-preserving Integration (PSI) tool. Study participants succeeded at assigned tasks, but also identified areas of confusion and error. Sarathy et al. [59] conducted a usability study with 19 non-expert participants using the DP Creator prototype to understand perceptions, challenges, and opportunities around DP analysis. Their findings highlight user challenges including users' poor understanding of decision implications, and difficulty accessing raw data and managing workflows. We expand prior research to evaluate the usability of multiple DP tools with data practitioners.

Recently, Govtech Singapore conducted a usability assessment of DP tools [61]. They compared the same four Python-based DP tools' capabilities in analysis, security, usability, and differential privacy, generating a usability benchmark for these tools. This was an expert heuristic review without user testing, which can be subjective and lacks depth compared to our usability study that involves data practitioners.

Usability of Non-DP Tools. While our study focuses on DP tools' usability, it is critical to draw implications from prior research on non-DP tools. There is usability research on non-DP data science tools that require programming skills. Akil et al. [4] compared the usability of three prominent distributed data processing platforms for cloud computing (MapReduce, Spark, and Flink). They found ease of use, learnability, language support, auto-configuration, and community support can make big data platforms more usable to data scientists. Mehta et al. [42] evaluated five large-scale image analysis systems (SciDB, Myria, Spark, Dask, and TensorFlow) and found various usability problems, including lack of support for user-provided Python code and manual tuning requirements for efficient execution. These studies show that data science tools often fail to support data practitioners in certain data processing and analysis tasks beyond programming. Since applying DP involves data science tasks, our study investigates if DP tools share similar usability issues as other data science tools.

Software engineering researchers have identified many usability issues in the technical documentation provided by developer- or programmer-facing software tools [2, 3, 43, 63], such as inconsistent content quality (e.g., readability, completeness, up-to-dateness) and poor navigation within the documentation. Additionally, Becker et al.'s systematic review of text-based programming error message research revealed diagnostic messages generated by compilers are often unhelpful to programmers [10]. For example, compiler error messages written in natural language were as difficult to read as source code [8], and many error messages were poorly designed, particularly for novice programmers [55]. Recommendations to improve programming error messages include increasing error message readability, reducing users' cognitive load, providing context to localize the problem, and showing examples, solutions, or hints to programmers [8, 9, 33, 62]. In our study, we also examine how DP tools could leverage existing usability best practices from these non-DP software tools.

3 Methods and Study Design

We chose usability testing methods [23, 50] to observe data practitioners' efforts to understand and implement DP using DP tools. Usability testing can identify impediments to data practitioners' DP implementation. We used surveys, interviews, and think-aloud protocol [18] for the data to answer our research questions. We executed the usability test remotely to reach a wider pool of participants. Research has shown that remote synchronous usability tests align closely in efficacy with traditional lab-based tests [5].

3.1 Selection of Differential Privacy Tools

To select tools for our study, we first conducted a review of available DP tools and decided on inclusion criteria based on study goals and feasibility that would allow for direct comparisons between tools. We required that tools: (1) be open source, (2) support standard statistical queries (count, sum, average, etc.), (3) have comprehensive documentation, and (4) provide a Python API. Based on these criteria, we did not include graphical applications like DPCreator [22] or Private data Sharing Interface (PSI) [27], or machine learning tools [53, 68]. We eliminated Chorus [36, 58], GoogleDP [30], Privacy on Beam [57], PINQ [41], and ZetaSQL [65, 69] due to lack of Python support. We included the remaining four tools in our study: OpenDP [52], PipelineDP [54], DiffPrivLib [21], and Tumult Analytics [11].

3.2 Study Procedures

3.2.1 Recruitment & Screening

This study received approval from our university's Institutional Review Board (IRB). We conducted a pilot study with four graduate students (one per tool) from our university and compensated them 25 US dollars each. Outcomes included adjusting study time allocation, increasing participant com-

ID	Tool	DP Expertise	DP Answers Correct	Black or Hispanic	Non-Male
E005	DiffPrivLib	Expert	4/4		
E008	DiffPrivLib	Expert	3/4	x	x
E011	DiffPrivLib	Expert	4/4		x
N002	DiffPrivLib	Novice	0/4		x
N004	DiffPrivLib	Novice†	3/4		x
N011	DiffPrivLib	Novice	1/4		x
E002	OpenDP	Expert	3/4		
E007	OpenDP	Expert	4/4		
E012	OpenDP	Expert	4/4		
N003	OpenDP	Novice	2/4	x	x
N008	OpenDP	Novice	2/4		x
N012	OpenDP	Novice†	3/4	x	x
E001	PipelineDP	Expert	4/4		x
E004	PipelineDP	Expert	4/4		
E009	PipelineDP	Expert	3/4		
N005	PipelineDP	Novice	1/4		x
N009	PipelineDP	Novice	1/4		x
N013	PipelineDP	Novice	1/4		
E003	Tumult	Expert	3/4		
E006	Tumult	Expert	4/4	x	
E010	Tumult	Expert	4/4		
N006	Tumult	Novice	1/4		x
N007	Tumult	Novice†	3/4		
N010	Tumult	Novice	0/4		x

Table 1: Summary of 24 study participants. The † symbol denotes participants who were initially categorized as DP experts by the eligibility survey but then re-categorized based on incorrect answers to DP questions in the post-task interview.

pensation, and clarifying survey and interview questions.

We aimed to recruit at least 24 *data practitioners*, with a balanced ratio between *DP novices* and *DP experts*, according to best practices for usability testing with developers in the privacy and security field [1]. We posted the study recruitment advertisement with a link to our eligibility survey on the Women in Machine Learning and OpenDP mailing lists, on Reddit in data science-related subreddits, and on LinkedIn.

The eligibility survey (Appendix A) determined participants' eligibility, obtained potential participants' informed consent, and gathered information about their data science and DP expertise. We deemed respondents eligible if they self-reported adequate data science experience (questions 1-3) and correctly answered at least one Python question (questions 4-5). We initially categorized respondents to be *DP experts* if they correctly answered 3 out of 4 DP knowledge questions (questions 8-11), and *DP novices* otherwise. We finalized the DP expert/novice categorization after each session by assessing participants' answers to DP questions in the post-task interview (Appendix D) since multiple-choice questions in the eligibility survey were subject to guessing. This led to the re-categorization of 3 participants as DP novices (see Table 1).

Of the 109 respondents who started our eligibility survey, 83 completed it and 47 were eligible. We invited all 47 eligible respondents to the study, prioritizing underrepresented females due to diversity goals and timeline constraints.

We chronologically assigned confirmed participants to tools equally using the initial DP expert/novice categorization. After the initial tool assignment, we confirmed with each participant that they had not used the assigned DP tool before. We continued recruitment after adjusting 3 participants' expert/novice categorization until we reached our recruitment target with a balanced expert/novice ratio.

26 confirmed participants completed their study sessions but we excluded two from data analysis (N001, E012): One due to the participant's inadequate Python skills, and the other due to an unavoidable session disruption that shortened task completion time. A summary of the 24 study participants, their tool assignments, and their responses to the eligibility survey appear in Table 1. Participants' ages spanned from 18 to 40, but most (14) fell between 25-34 years. Our sample consisted of 54% females, 38% males, 4% non-binary individuals, and 4% who chose not to specify their gender. We conducted all usability test sessions on Microsoft Teams, following specific guidelines to maintain consistency. After the study session, each participant was compensated with a gift card of 40 US dollars for up to 1.5 hours of study time.

3.2.2 Pre-task Procedures

Before commencing usability study tasks, we made sure participants shared their screens and understood the think-aloud protocol. Participants also reviewed a handout that covered the fundamentals of DP and a tutorial of their assigned DP tool with executable sample DP tasks in Jupyter Notebook (see Appendix B). The handout and the tutorial provided participants necessary background for the study tasks but may introduce confounding factors to study results (detailed in Section 5.1).

We informed participants that they could refer back to the handout and the tutorial, consult the tool's official documentation, and use Google search during the study. We asked them not to use how-to resources, like StackOverflow. This ensured that participants had access to essential general resources (e.g., Python libraries) to complete the study tasks, but not to existing solutions to prevent cheating.

3.2.3 Usability Testing Tasks

We designed three usability testing tasks on differentially private data analysis, shown in Table 2. We modeled the tasks on a demo in Pipeline DP's documentation [54], changing it to a new, synthetic dataset that counted restaurant visits across a week (see Appendix E). Our tasks involved common data analysis operations supported by all four DP tools (i.e., count, sum, mean). Participants had one hour to complete the tasks by writing Python code in a shared Jupyter notebook.

The three tasks were the same across DP tools. The assigned total privacy budget was $\varepsilon = 1.2$ for all the tasks. Participants could set all other parameters themselves (including the per-task privacy budget). We encouraged participants to articulate their thought process using the think-aloud method, and we recorded both their spoken insights and on-screen

Task	Description
Task 1	How crowded is the restaurant on weekdays? (total number of visits for each weekday)
Task 2	Total amount of time spent by visitors on each weekday (exclude weekends).
Task 3	Average amount of time spent by visitors on each weekday (exclude weekends)

Table 2: Usability testing tasks. See Appendix E for details of the dataset used and Appendix F for solutions.

actions during the study.

3.2.4 Post-task Procedures

Participants completed a post-task survey and a post-task interview (Appendices C and D). The survey repeated DP questions from the eligibility survey to assess participants' DP understanding after the study. It also gathered data on participants' study experiences. The post-task interview gathered qualitative data for deep insights into participants' challenges during the study and their preferences for DP tools.

3.3 Usability Measurements and Data Analysis

3.3.1 RQ1: DP Understanding

Even experienced data scientists sometimes fail to grasp the intricacies of DP [19, 67]. The DP tools in our study all aim to make DP more understandable to data practitioners. To assess these tools' effectiveness in supporting DP understanding, we used the same four DP knowledge questions in the eligibility survey and the post-task survey to compare participants' pre-task and post-task DP knowledge differences. To mitigate confounding factors introduced by pre-task procedures, we also analyzed participants' explanations of key DP concepts in our post-task interview and their reported useful sources for DP understanding from post-task survey and interview.

3.3.2 RQ2: DP Implementation

We used three widely-used usability criteria — learnability, efficiency, and error prevention — to assess how effective the tools support DP implementation [51]. **Learnability** measures if new users can successfully use a specific tool or interface. We use task success and failure rates [12] to measure DP tools' general learnability. Specifically, we evaluated whether users succeeded or failed to complete tasks and assessed the correctness of their completed tasks against our reference solutions. **Efficiency** measures how fast users can accomplish tasks with a specific tool or interface. We recorded the time taken to complete each task to measure DP tools' efficiency. **Error prevention** is about how well a tool prevents user errors and, in the cases of error, how well a tool facilitates error identification and recovery. We define errors during DP implementation as interruptions of progress toward task completion and qualitatively analyzed these interruptions from the screen recordings, think-aloud, and post-task interviews. Additionally, we analyzed participants' post-task survey responses to

identify the factors that impacted DP implementation.

3.3.3 RQ3: User Satisfaction

We first quantitatively evaluate user satisfaction using the two standardized measurements: the System Usability Scale (SUS) [13] and the Net Promoter Score (NPS) [31]. Since DP tools are specialized data science tools, we slightly customized the wording of SUS and NPS. We also analyze the qualitative data from post-task interviews, including their overall user experiences and areas of improvement, to yield insights into user satisfaction with these DP tools.

3.3.4 Mixed-Methods Data Analysis

We report the descriptive statistics by tool to allow usability comparison across the four DP tools examined. We also report key statistics by participants' prior DP expertise level, either expert or novice, so that usability is recognized relative to participants' prior DP knowledge. Due to the small sample size, we refrained from performing statistical tests to avoid over-generalization (details in Section 5.1).

The first and the second authors also rigorously analyzed the qualitative data collected from this study, including transcripts of audio recordings, video recordings of participants' screens, and Jupyter notebooks from all sessions. The two authors used a hybrid thematic analysis approach combining inductive and deductive coding [26] to annotate the data. They created the initial codebook from the pilot sessions and continuously refined it through research team discussions during the full study data analysis. The finalized codebook included both qualitative codes (e.g., type of challenges during implementation, misunderstandings of DP concepts) and quantitative counts derived from qualitative assessment (e.g., number of correctly completed tasks, time taken for each task). Then the first and the second authors independently coded all qualitative data using the codebook. They resolved all coding conflicts either through their own discussion or through seeking consensus from the whole research team.

4   Results

4.1   RQ1: DP Understanding

4.1.1   Pre- and Post-Task Response to DP Questions

Figure 1 reports the number of correct answers to the DP knowledge questions before and after study tasks, organized by participants' DP expertise level and by tool. Specifically, Figure 1a shows that experts provided similarly high-level of correct answers pre- and post- tasks, possibly due to their familiarity with DP concepts. However, novices showed a boost in their DP understanding as shown by the rise in correct answers from pre-task to post-task. This result indicates that our study procedures, including the DP implementation tasks, particularly helped novices understand DP concepts. Figure 1b shows the pre- and post-task DP knowledge difference across tools. All of the tools except OpenDP boosted participants' DP understanding, where DiffPrivLib saw the

greatest jump from 15 to 20 correct answers. Note that the study-provided handout and tutorials also impact participants' post-task DP understanding (see Section 4.1.3).

4.1.2   Participants' Explanation of DP Concepts

To further investigate participants' understanding of DP, we looked at how they described key DP concepts in their own words during the post-task interview (Appendix D questions 2-3), focusing on DP, ε, privacy budget for each task, and total privacy budget for all tasks. We aimed to see if participants understood that the privacy budget and ε essentially refer to the same concept and that the total privacy budget across multiple analyses accumulates the ε values (i.e., sequential composition). We considered participant responses to be correct if they were factual and includes details similar to our sample correct answers in Appendix D.

Table 3 details the number of participants who could accurately explain DP concepts, divided both by their level of expertise and by the assigned tool. All 12 expert participants accurately explained the concept of DP. For example, one expert provided a robust definition, stating, *"Differential privacy is a mathematical definition for privacy that basically says that if we compute an analysis with a particular individual's data or without it, we should get similar outputs. Whether or not somebody participates in the data set, the outcome should be pretty much the same"*(E011). This explanation is correct because it clearly describes how randomness is used in analyses to ensure results are consistent, whether an individual's data is included or not, and emphasizes the importance of the privacy parameter. When discussing the privacy budget, 11 out of 12 experts explained how the budget was allocated in individual tasks, and 10 out of 12 were able to describe how these budgets add up to form the total privacy budget.

Only 9 out of 12 novices could adequately explain DP in their own words, often missing critical details. A typical novice explanation was less precise, *"From what I remember, it's like some sort of tool-based guarantee for privacy over millions of users..."* (N006), which lacks specificity and critical details about the mechanism of DP, like the privacy parameter. In their understanding of the privacy budget for each task, 8 out of 12 novices had a basic grasp. 7 out of 12 demonstrated an understanding of privacy budget accumulation, suggesting areas of confusion among novices.

The above difference between experts and novices shows the importance of participants' prior DP knowledge on their understanding. Additionally, each assigned tool had no clear impact on participants' understanding of DP, as reflected in Table 3.

Notably, participants gave incorrect answers for the question "what was the total privacy budget for the whole notebook?" more often than for the other questions (Table 3). This result was consistent across experts and novices, and across tools, suggesting that composition is a difficult concept and should be made clear by DP tools. For example, one novice
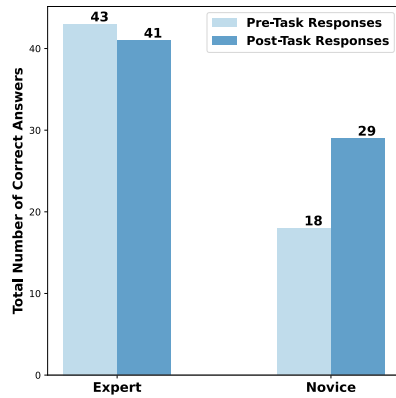
(a) By expertise level
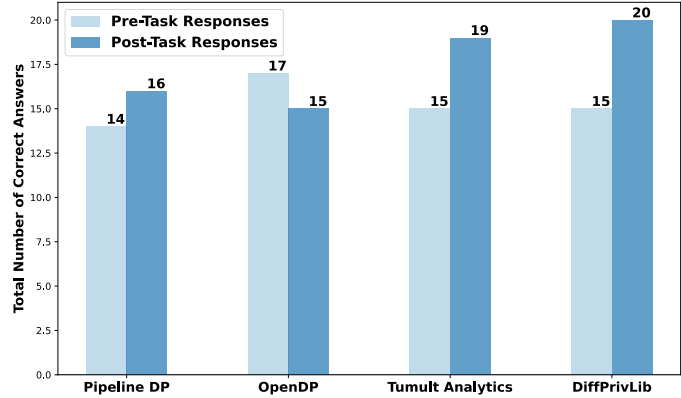
(b) By tool

Figure 1: Total number of correct answers to DP knowledge questions before and after study tasks.

participant answered: *"I got confused between these, total budget and the amount of epsilon for each of the individual tasks. That part, I didn't get it<"* (N010)

4.1.3 Useful Sources for DP Understanding

Participants selected all sources that helped them understand DP concepts during the study in the post-task survey, as shown in Figure 2. The figure displays the average rankings of resources, where resources are ranked based on participants' preferences from 1 (most helpful) to 4 (least helpful). Figure 2a indicates that the handout and the tutorials supported their DP understanding more than DP tools' official documentation across all tools, while participants' prior DP knowledge played a key role. Figure 2b shows that experts relied heavily on their prior DP knowledge, while novices used the handout and the tutorials to understand DP concepts. This result suggests educational sources like the handout and the tutorials provided in this study benefit data practitioners' DP understanding, while DP tools' documentation lacks such support.

Post-task interview data suggests that concrete examples (like the ones in our tutorials) and short explainers (like the handout) helped participants understand important DP concepts, as E001 commented: *"It also helped to have the tutorial. ... if you had only given me the documentation... it would have taken me much longer to put it together."* Note that the handout and the tutorial were part of the study instrument, so we cannot fully attribute novices' increased DP understanding in Figure 2a to the DP tools themselves.

Moreover, our qualitative data indicates that participants could use more help with DP's actual privacy protection. In the case of ε-values and privacy budgets, we asked participants how strong they thought the privacy protection was for their just-completed task. One DP expert (E006) confidently said:*"That's the hard question to answer. The total privacy budget for all of the tasks was 1.2, a value that is in line with recommended guidelines. [ε] is around 1.0. So, maybe that's somewhat strong."*. Other responses lacked

consistency and confidence: *"I think [ε] should be much lower...probably around .5 or probably even lower..."*(E003) and *"Pretty strong...very strong, actually"*(N007).

4.2 RQ2: DP Implementation

4.2.1 Learnability

Task completion and correct rates. To measure learnability, we evaluated the completeness and correctness of participants' solutions. We considered tasks **complete** when code executed without error and produced correctly formatted responses, and **correct** when they satisfied DP and had comparable utility to our reference solutions.

Figure 3a shows the completion rates for three usability testing tasks across four tools: all DiffPrivLib participants completed all three tasks, while none of the OpenDP participants completed tasks #2 or #3. Tumult Analytics and PipelineDP results fall between these two extremes, with all participants completing at least task #1.

The varying completion rates may derive from the different API designs of the tools. DiffPrivLib provides a minimal API and encourages users to use it in combination with Python data analytics libraries like Pandas. Similarly, Tumult Analytics mimics an existing data analytics API called Spark. OpenDP, in contrast, does not leverage mainstream Python libraries for a learning scaffold. Participant comments on API design from post-task interviews lend support to this finding. For example, one expert (E006) liked the similarity of the Tumult Analytics to Spark: *"I think the fact that it was very similar to Spark was really helpful...I have a decent amount of experience with Spark and Pandas, so that was very intuitive."*

Figure 3b shows the task correctness rates. Some participants completed tasks but incorrectly, so the correctness rates are no larger than the corresponding completion rates. Combined, the completion and correctness rates show that: (1) complete Tumult Analytics and OpenDP solutions were all correct; (2) complete PipelineDP solutions were mostly—but

Question	Experts n = 12	Novices n = 12	DiffPrivLib n = 6	OpenDP n = 6	PipelineDP n = 6	Tumult n = 6
After completing the tasks, can you explain differential privacy to me in your own words?	12	9	5	6	5	5
What was the privacy budget for each task?	11	8	4	5	5	5
What was Epsilon?	12	9	5	6	5	5
What was the total privacy budget for the whole notebook?	10	7	4	4	4	5

Table 3: Number of correct answers to post-task survey questions measuring the understanding of DP concepts, disaggregated by level of expertise and by tool. Experts answered more of these questions correctly than novices, but the assigned tool had no clear impact on the number of correct answers. See Appendix C for sample correct answers.

(a) Useful sources by tool

(b) Useful sources by expertise level
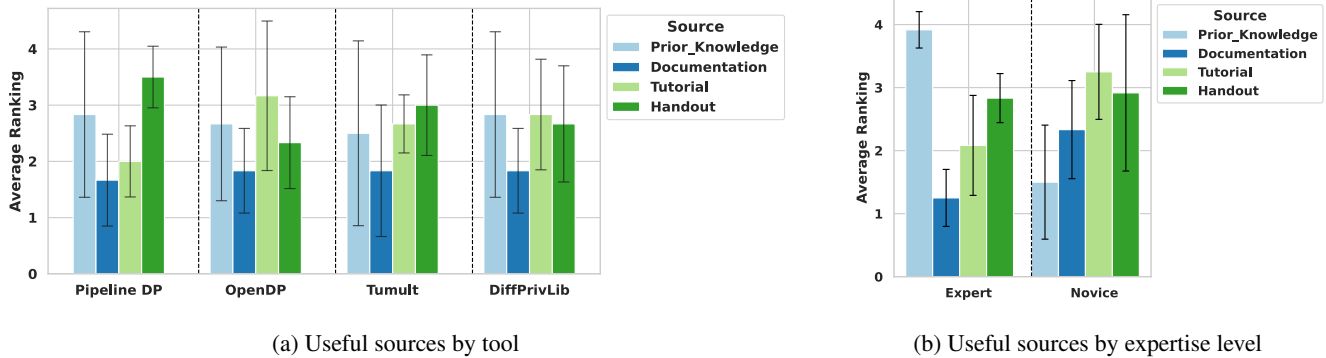
Figure 2: Useful sources that support participants' DP understanding, by tool (a) and by expertise level (b).

(a) Task completion rates by task and tool

(b) Task correctness rates by task and tool
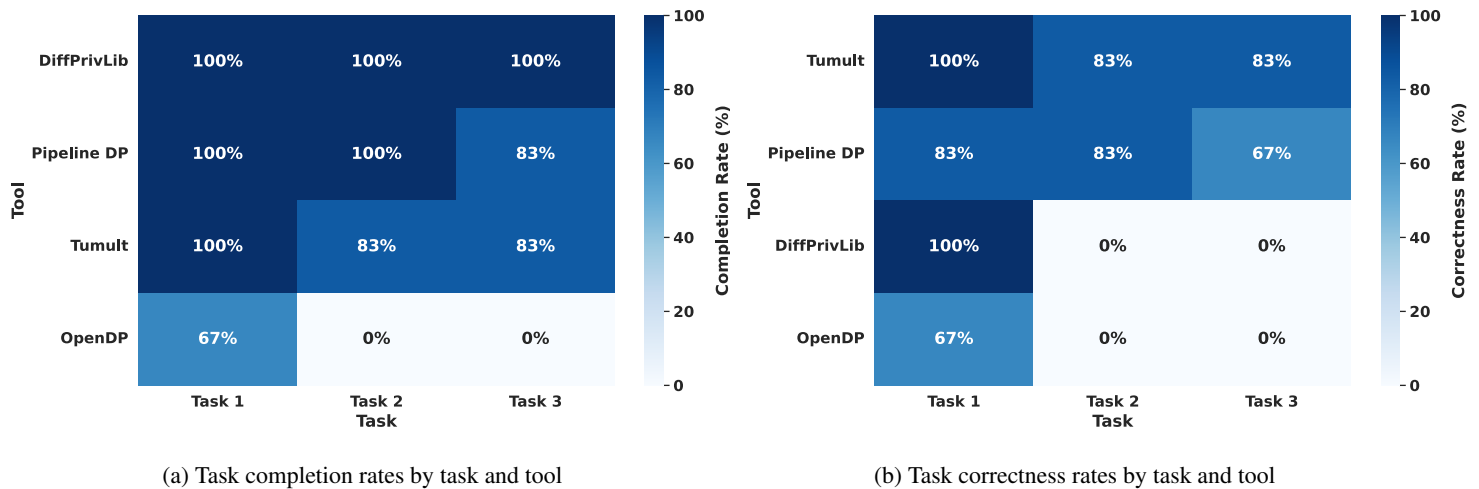
Figure 3: Learnability of DP tools measured by (a) task completion rates and (b) task correctness rates. Each cell represents the percentage of participants who completed or correctly completed the task using the tool.
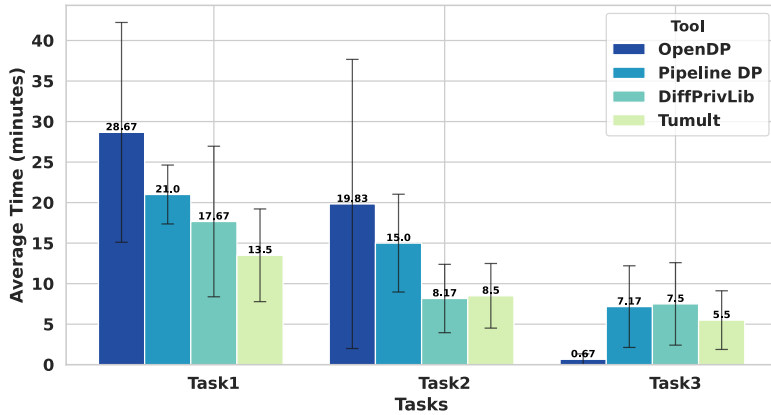
(a) Average time taken by tool
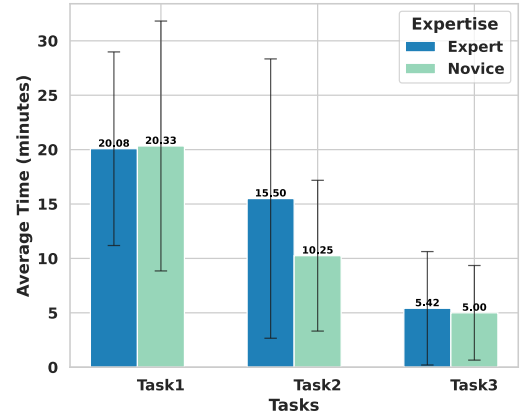
(b) Average time taken by expertise level

Figure 4: Average task completion time: (a) by tool (b) by expertise level.

not all—correct; (3) Complete DiffPrivLib solutions were all *incorrect* for tasks #2 and #3.

Causes for incorrect implementation. Qualitative analysis of the screen recordings revealed the causes of some incorrectly completed tasks First, all six DiffPrivLib participants failed to apply the correct **sensitivity**, which refers to the upper and lower bounds that provide the extent of valid DP, in tasks #2 and #3. (Task #1, a counting query, has a sensitivity of one, a value that is intuitively correct.) DiffPrivLib does not signal any error related to sensitivity bounds, even though this mistake violates DP. Some expert participants were uneasy about their approach for setting sensitivity, but even these participants were not able to produce correct solutions.

Second, some tools lack **feedback** about query results' correctness. For example, one PipelineDP participant (E004) used strings (rather than integers) as grouping keys, resulting in histograms containing only 0s, and the participant did not notice the mistake. The participant later discussed this in the post-task interview, *"It's the right number of attributes and it's the right metric...the result is very noisy,"* but he added, *"I don't know if there's a way to check the final [privacy] budget."* In this case, Pipeline DP's lack of feedback affected the solution's correctness but did not violate DP.

Finally, confusion about **whether and how the tools handle the privacy budget** led to incorrectness, particularly for Pipeline DP and DiffPrivLib. E009 commented on PipelineDP: *"I would expect maybe that [a] budget accountant object could tell me my budget so far. [I'm] looking for a way to figure out how much I spent so far."* And N011 on DiffPrivLib: *"[I'm] confused about how the privacy budget would be handled at the object level. When creating the mechanism objects, should I use the same object for every analysis...and the ε will add up to the right number...can you compose all of those together? That wasn't totally clear to me."*

4.2.2 Efficiency

To measure efficiency, we calculated the time taken to complete each task by reviewing the screen recordings.

Figure 4a shows the time taken on each task by tools. OpenDP participants spent the most time on task #1 (nearly 30 minutes on average), while Tumult Analytics participants spent the least (fewer than 15 minutes on average), with DiffPrivLib and PipelineDP falling in between. The time taken for Task #2 shares a similar trend while all participants spent less time on task #2 than task #1. However, time taken for task #3 varied. OpenDP participants spent almost no time on task #3, while participants using the other three tools spent similar amounts of time on task #3, but less than that of tasks #1 and #2. The total time limit (1 hour) imposed on all tasks may affect the time spent on task #3. OpenDP participants spent nearly all of the allotted time on tasks #1 and #2, leaving little time for task #3. Participants using the other tools either finished task #3 quickly or ran out of time. *"I think I wasted a lot of time trying to find what I don't know,"* said E013.

Figure 4b shows the time spent on each task, by participants' expertise level. For tasks #1 and #3, novices and experts took roughly the same amount of time; for task #2, however, experts took *longer* than novices. Qualitative analysis from participants' think-aloud showed that experts' confidence, curiosity, and skills prompted them to explore task solutions. E005 spent time *"investigating the number of visitors that show up multiple times per day"* only to find the occurrences are rare in the data, concluding that *"we can just set the sensitivity to one."* Other experts also spent time examining API functions, honing DP parameters, checking results, and exploring alternative approaches. Novice users, in contrast, typically accepted the tool's default settings and did not spend time considering these issues. *"I'm not familiar with all the different functions,"* N011 told us in think-aloud while looking at different options for DP noise, and added post-task, *"The land of functions are [sic] totally wild to me."*

Tool	Stucks	Unstucks	Unstuck %
DiffPrivLib	31	27	87.1%
OpenDP	79	22	27.8%
Pipeline DP	54	38	70.4%
Tumult	43	38	88.4%

DP Expertise	Stucks	Unstucks	Unstuck %
Experts	105	64	63.8%
Novices	43	16	56.9%

Stuck Type	Stucks	Unstucks	Unstuck %
DP	4	3	75.0%
Documentation	65	31	47.7%
Python	28	27	96.4%
Results	18	4	22.2%
Task	34	31	91.2%
Tool	58	29	50.0%

Table 4: Stuck counts, unstuck counts, and unstuck percentages by DP tool, participants' DP expertise, and stuck type

4.2.3 Error prevention

We consider interruptions of progress toward task completion as errors during DP implementation and call them "stucks". We also examine error recovery when participants resolve these interruptions and call them "unstucks."

Stuck and unstuck statistics. We report the counts for stuck and unstuck, as well as the unstuck percentages in Table 4, organized by DP tool, by participants' DP expertise, and by stuck type (defined in Table 5). Tool-wise, participants assigned to Tumult Analytics (38/43, 88%), DiffPrivLib (27/31, 87%), and PipelineDP (38/54, 70%) often managed to get unstuck, but those assigned to OpenDP rarely got unstuck (22/79, 28%). Expertise-wise, DP experts (67/105, 64%) and novices (58/102, 57%) had similar unstuck percentages.

Stuck types. We identified six types of stuck in Table 5 and contextualized them with qualitative data. The most frequent stuck type was **documentation stucks** (65 counts, 48% unstuck percentage), where participants had difficulty finding answers in tools' documentation. *"I can imagine how to do this without this library,"* said E011 (DiffPrivLib), *"I'm trying to see...how to translate that into the library."* Second was **tool stucks** (58 counts, 50% unstuck percentage), where participants struggled to execute tools' function calls or to interpret tools' error messages. Participants would either fail to grasp tool basics: *"I don't get the terminology or the syntax,"* said E003 (Tumult Analytics); or, the issue with the tool was a specific aspect of DP: *"I'm trying to figure out how I actually tell the session what the sensitivity of the query is,"* said E006 (Tumult Analytics). **Task stucks** were common (34 counts, 91% unstuck percentage) but most participants got unstuck by asking researchers for task clarification. For example, days in our dataset are integers, 1-7. E005 (DiffPrivLib) asked, *"Can I ask is day one equal to Monday and Day 7 equal to Sunday?"* Participants also experienced **Python stucks**

(28 counts, 96% unstuck percentage) but almost always got unstuck by consulting Python sources.

Usability issues. Our qualitative analysis articulated how DP tools' usability issues with their documentation and APIs caused errors and hindered error recovery. **DP tools' documentation** presented many usability problems. E001 found the upper bound for data values in PipelineDP unclear: *"I'm not super sure about this maximum value because I'm not sure if I interpret it correctly [in] the documentation."* Other participants hoped the documentation would provide more details about different API functions, such as the best DP mechanism for a data analysis task. E005 commented on DiffPrivLib's documentation, *"I do think that sometimes when you present people with a suite of 16 options, it's important to detail what the differences are and when one option might be more effective than another."* The format of the documentation was also challenging. Participants struggled due to the lack of organization of OpenDP's documentation. E013, for one, *"got lost in it."* **DP tools' APIs** also caused DP-specific errors. Participants struggled with API instructions to set parameters for DP mechanisms. And if the tools' parameters were idiosyncratic, the user interaction was less intuitive. *"I was not confident because I didn't know what the library was doing"* and *"I wasn't sure what the argument [meant],"* E004 (PipelineDP) said, *"I don't really know in the end if I computed what I was really expecting to compute."*

Error recovery was challenging. N012 was frustrated by the lack of examples *"...I couldn't get examples of people running into the same problem."* Error messages sometimes were unhelpful. For example, OpenDP's API returned error messages in Rust, and not translated to the API's Python wrapper. E002 said: *"I don't really know any Rust. Coming from a Python experience, [it] might be better to have error messages in Python that indicate the error in the line of Python."*

4.2.4 Factors impacting DP Implementation

Participants' post-task survey responses revealed the factors that helped or hindered their DP implementation in the study. Full results appear in Figures 5 and 6 in Appendix G.

9 out of 12 novices and all 12 experts reported the tutorial helped their DP implementation, with tool documentation (5 novices and 8 experts) and their data science skills (7 novices and 8 experts) close behind. Notably, none of the participants assigned to OpenDP reported that their data science skills or the tool's documentation were helpful, possibly due to how OpenDP's API differs from mainstream Python libraries. E002, someone familiar with data frames and method chaining in other Python libraries, failed to understand basic OpenDP syntax, *"I don't know what you call that little stream operator thingy."* E012 said that *"it's written in a very C-heavy style as opposed to a Python-style that most people are used to."*

8 out of 12 novices were hindered by lack of prior DP knowledge, while 5 out of 12 experts reported being hindered by DP tools' documentation in completing the tasks. Novices

Stuck Type (Abbreviation)	Definition
DP misunderstanding (DP)	Incorrectly interpreting or applying DP.
Documentation stuck (Documentation)	Struggle to interpret documentation descriptions.
Expected result stuck (Result)	Answer from a DP tool query that is not in line with expected DP values.
Python stuck (Python)	Don't know the correct Python or Pandas function to use.
Question stuck (Task)	Misinterpretation of a Task assignment, or need to clarify a Task detail.
Tool stuck (Tool)	Don't know the correct DP tool function to use. Failing to interpret error codes.

Table 5: Definitions of six types of stuck from our qualitative analysis

like N009 (Pipeline DP) *"took a lot of time understanding what the metrics are and what each parameter is"* and N010 (Tumult Analytics) *"got confused between the total [privacy] budget and the [epsilon] for each of the individual tasks."* One expert, E003, asked for *"a step-by-step guide on how you can how you can use Tumult Analytics for your particular use case."* These results suggest that DP tools should help enhance novices' DP knowledge and improve their documentation to support experts' DP implementation.

4.3 RQ3: User Satisfaction

4.3.1 Quantitative Ratings

The Net Promoter Score (NPS) and System Usability Scale (SUS) metrics from the post-task survey showed that participants were most satisfied with DiffPrivLib and least satisfied with OpenDP. DiffPrivLib had the highest NPS (33.33), followed by Tumult Analytics (-16.67), PipelineDP (-33.33), and OpenDP(-66.67). Similarly, DiffPrivLib had the highest SUS score (63.89), followed by Tumult Analytics (57.64), PipelineDP (54.51), and OpenDP (38.19). Full statistics appear in Figure 7 in Appendix G. These ratings align with the task completion rates associated with each tool (Figure 3a)—DiffPrivLib had the highest user satisfaction ratings and the highest completion rate, followed by Tumult Analytics, PipelineDP, and OpenDP. This suggests that participants were most satisfied with tools that made it easy for them to complete the study tasks.

4.3.2 Qualitative Results by Tool

Our qualitative results from the post-task interviews triangulated the above quantitative ratings and articulated participants' user experience with each tool, as described below.

DiffPrivLib received positive comments about its API and documentation: *"I liked the API of the tool. I thought the documentation was pretty clear" (E005)*. Participants also liked its compatibility with familiar libraries: *"I really liked that it integrated nicely into a library that I already have worked with, Pandas..." (E011)* and felt comfortable with the tool by the end of the session: *"Now...I'm on task three, I feel like I have a hang of the pattern...this isn't adding that much more time to my typical process" (E011)*.

Tumult Analytics was acclaimed for its intuitive API, as E010 said:*"Similarity with Pandas was definitely a plus. That's probably the best thing they've done there.*" However, E003

expressed frustration with its documentation: *"It was just a single-page documentation and I had to like scroll all the way down to find the exact syntax."*. The feedback addressed the user need for improved documentation navigation.

PipelineDP exhibited documentation challenges: *"The documentation was quite incomplete...sometimes it just had one sentence about terms like 'Max contribution' or 'Max value' and it wasn't really clear to me what that meant" (E004)*. Participants also wanted the ability to search: *"What [does] the documentation say about the budget? I don't have a way to search this page" (E001)* and found error messages confusing: *"I think the error message wasn't super clear and it would be tough to debug" (E004)*. Several participants wished for examples in the documentation: *"Functions should contain some examples...[like] what each parameter is..." (N009)*.

OpenDP had usability issues with its error messages and documentation: *"The error messages I'm getting here come from Rust and I don't know what it means" (E007)*, and *"The documentation wasn't useful...[I] felt like it was a little confusing...." (N008)*. OpenDP participants also wished for examples:*"It would have been a lot more helpful if there were examples" (N012)*.

Overall, these findings on user satisfaction echo prior results — DP tools' API design and documentation quality are paramount to data practitioners' user experience.

5 Discussion and Recommendations

5.1 Limitations

We acknowledge several limitations of this study. First, we only evaluated four open-source Python-based DP tools for fair comparison across tools so that our results cannot represent all DP tools. Similarly, the findings may not generalize to all data practitioners due to our small US sample. However, our sample is similar to prior usability studies evaluating security/privacy tools with developers [39, 48] to generate valid insights. Therefore, we refrained from performing statistical tests to avoid over-generalization of the statistical results from this study to all DP tools or data practitioners. Instead, we emphasized key descriptive statistics and qualitative results.

Second, our study instrument introduced confounding factors because the handout and tutorials (Section 3.2) helped participants understand DP and complete study tasks. However, we had to prioritize study feasibility to ensure participants

with varying prior knowledge had the necessary information to get started. To minimize this bias, we ensured our handout and tutorials did not reveal answers to study questions or tasks, and we remained cognizant when analyzing and reporting study results.(see Sections 4.1.3 and 4.2.4).

Moreover, we only evaluated the usability of three first-step DP data analysis tasks. The results may not reflect the usability of the full capability of the examined DP tools. However, usability issues surfaced in these first-step tasks hinder developers' adoption of software tools or APIs [1,47], our recommendations for usability improvements still benefit other DP tools and encourage overall DP adoption.

5.2 Provide Usable Documentation

Our results highlighted usability issues with DP tools' official documentation, leading to the following recommendations.

Improve documentation navigation. Participants generally experienced difficulty navigating DP tools' official documentation, including technical documentation on APIs. Firstly, despite the fact that all four tools provide how-to guides with code examples in their documentation, participants struggled to find specific guides that matched their data analysis tasks at hand. For example, the descriptions of these guides are often generic and contain DP terminology (e.g., "how to perform counting queries with the Laplace mechanism"), which is unfriendly to DP novices. The mismatch between documentation contents and the practical development tasks often caused poor documentation findability [3], which can be mitigated by providing more accurate and readable task descriptions that align with users' goals [63]. Secondly, our participants disliked DP tools' single-page formatting (see Section 4.3.2). This formatting uses a single web page to organize documentation for every API function within a module, which can be lengthy, worsening the findability problem. In contrast, mainstream Python libraries (e.g., NumPy, Pandas) use one page per documented function and are easier to navigate. Additionally, some participants also hoped to be able to search within DP tools' documentation, which also resonates the proposed techniques to improve the usefulness of software documentation [2]. Our findings echo prior software engineering research on usable documentation [2,3]. We believe DP tools can leverage existing best practices for good software documentation, such as providing intuitive task descriptions, improving information organization, and adding a search or recommender tool to improve documentation navigation.

Include DP-specific examples and advice. Many participants found the documentation for the API function they wanted to use but had trouble understanding the descriptions of DP-specific parameters and were not able to find examples that made use of the documented function. Some requested more use cases and code examples within DP tools' documentation (see Section 4.2.3). These results are consistent with prior research on developers' need for documentation [2,43].

Moreover, some DP novice participants had trouble deciding which DP mechanism to use—for example, when given a choice between the Laplace, Gaussian, or Geometric mechanisms. Existing tool documentation fails to address these questions since it predominantly emphasizes **how** to use a specific mechanism rather than **which** mechanism to choose. To make DP tools' documentation truly usable for data practitioners, we recommend that DP tools go beyond generic best practices for usable documentation and include DP-specific advice that would particularly benefit DP novices.

5.3 Improve Error Prevention & Recovery

The study findings yielded rich insights into how DP tools can prevent errors and help users recover from errors.

Warn users about severe DP violations. PipelineDP, Tumult Analytics, and OpenDP actively prevent DP violations—they require users to wrap sensitive data using special objects. They provide error messages when users attempt to perform actions that would violate DP. DiffPrivLib, on the other hand, relies on the user to avoid DP violations; for example, DiffPrivLib asks users to set the sensitivity for every mechanism and does not check that the specified sensitivity has been correctly enforced for the input data. This explains that all of the participants assigned to DiffPrivLib completed all three tasks, but *every single participant* violated DP in their solutions for tasks #2 and #3 and failed to correctly complete them (see Figure 3 in Section 4.2.1). Thus, we recommend that DP tools proactively warn users when DP might be violated.[1]

Improve error messages. When errors occurred, many participants had difficulty diagnosing and recovering due to poorly designed error messages (Sections 4.2.3 and 4.3.2). In particular, participants assigned to PipelineDP and OpenDP described confusion over the meaning of error messages, and trouble finding documentation to understand and fix the problem. This resonates with prior research on unhelpful compiler error messages of non-DP tools [8,9,55].Additionally, OpenDP further confused users who primarily have a Python background with error messages generated in the programming language Rust. We first recommend DP tools learning from general best practices to improve error message readability and provide examples, solutions, and hints [10]. Additionally, DP tools should consider the average DP knowledge of their intended users and offer support when the error is DP-related (e.g., pointers to resources on DP violations).

Ensure clarity in privacy budget setting and tracking. Some participants failed to explain the total budget (Section 4.1.2) and many were concerned with setting or tracking the privacy budget with different DP tools. Tumult Analytics

---

[1]DiffPrivLib raises a "privacy leakage warning" in some situations that may violate DP (e.g., when setting parameters based on the data), but not in all such cases. In particular, when the programmer uses an external library like Pandas to produce an aggregate result—as all of the participants in our study did—DiffPrivLib cannot enforce sensitivity bounds on the query and does not raise a warning.

asks users to set the total and per-query budget with required API calls. This process was not as clear in other DP tools: Some participants assigned to PipelineDP and DiffPrivLib were not sure whether the library keeps track of the privacy budget at all. This confusion did not necessarily result in failure to complete the study tasks, but it would result in unintended DP violations in real-world implementations. We recommend that DP tools clearly convey how to set the privacy budget and how the tool accounts for the total budget.

Balance DP violation prevention and general usability. We also observe the tension between preventing DP violation errors and maintaining the tool's usability (Sections 4.2.1 and 4.3.1). OpenDP's strict API was effective at preventing DP violations, but OpenDP had lower completion rates and satisfaction ratings. DiffPrivLib's flexible API resulted in many DP violations but received high completion rates and satisfaction scores. Tumult Analytics seems to strike the best balance. Its API was effective at preventing DP violations where users had high completion rates and satisfaction ratings. This indicates that DP tools may need to balance between their goal to prevent DP violation errors and the tool's usability.

5.4   Make API Design Intuitive

Our findings reveal participants' unique experiences with the APIs of DP tools, leading to the following recommendations.

Leverage users' familiarity with mainstream APIs. Results in Sections 4.2.1 and 4.2.2 suggest that participants implemented DP more successfully with DP tools that incorporate mainstream APIs that they are familiar with. Specifically, the intersection of DiffPrivLib with ubiquitous libraries like Pandas, garnered commendation. This cohesive integration provided a scaffold for new learning and obviated the need for relearning. Tumult Analytics was also appreciated for the way its API mimicked that of Spark. In contrast, PipelineDP provides an API centered on performing multiple aggregations at once, and OpenDP provides an API that focuses on transformations and composition. Neither is similar to mainstream data science APIs, which impeded participants' DP implementation in the study. To make APIs more usable, we recommend DP tools prioritize API designs that allow data practitioners to transpose their extant data science knowledge to the DP context, augmenting overall satisfaction.

Assist users in setting DP-related metadata via APIs. Setting DP-related metadata (e.g. total privacy budget, ε per query, upper bound on data values) is key to DP implementation. DiffPrivLib includes **default values** for metadata. The choice to use default values simplifies the API, but may result in users accidentally accepting inappropriate default values. DiffPrivLib provides warnings when default values could result in DP violations. This helped participants to complete the tasks correctly and suggests that default values can be effective if appropriately selected and implemented. The other three tools require **users to specify DP-related metadata** via

APIs. Experts appreciated that Tumult Analytics explicitly asks users to set per-query and total privacy budgets. However, DP novices may struggle with manually setting metadata, like with other non-DP tools [42]. Participants found PipelineDP's API for setting metadata confusing and struggled with settings like `max_value`, `partition_extractor`, and `privacy_id_extractor`. For OpenDP, our participants found its API, including the metadata portion, difficult to use.

We recommend that DP tools should**carefully design APIs to obtain this metadata**, as well as assist users in configuring key DP-related metadata, including exposing metadata settings, providing documentation for each metadata setting, and auto-filling appropriate default values.

5.5   Help Users with DP Foundations

Our study surfaced a general need for additional resources to help data practitioners better understand DP concepts. We found that many novices had difficulty understanding and describing the privacy budget (Section 4.1.2), and that both novices and experts sometimes had trouble describing the strength of the privacy guarantee (Section 4.3.2). These results reinforced previous findings that DP concepts are complex and difficult to communicate [14, 19, 38, 66], which inspire the following recommendations to address the challenge.

Provide general educational materials. Section 4.1.3 suggests that our study instrument boosted participants' DP understanding. DP tools may be able to replicate this effect by providing or directing users to general DP educational materials, similar to the handout and tutorials in this study.

Support privacy guarantee communication. Our participants had difficulty explaining the strength of the privacy guarantees, and several participants were unsure if their DP outputs would be private enough to be shared or published. We encourage DP tools to provide users additional community resources [4] on privacy guarantee (e.g., how to communicate the guarantee when disseminating DP analyses.)

6   Conclusion

We presented the first comprehensive usability study that evaluates four open-source Python-based DP tools with data practitioners. Our findings suggest that DP tools should provide easy-to-navigate, DP-specific documentation, enhance error prevention and recovery capabilities, improve API designs to ease users' learning curves, and offer resources to strengthen users' DP foundations. We aim for our findings and recommendations to facilitate broader DP adoption.

Acknowledgments

References

[1] Yasemin Acar, Sascha Fahl, and Michelle L Mazurek. You are not your developer, either: A research agenda for usable security and privacy research beyond end users. *2016 IEEE Cybersecurity Development (SecDev)*, pages 3–8, 2016.

[2] Emad Aghajani, Csaba Nagy, Mario Linares-Vásquez, Laura Moreno, Gabriele Bavota, Michele Lanza, and David C Shepherd. Software documentation: the practitioners' perspective. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*, pages 590–601, 2020.

[3] Emad Aghajani, Csaba Nagy, Olga Lucero Vega-Márquez, Mario Linares-Vásquez, Laura Moreno, Gabriele Bavota, and Michele Lanza. Software documentation issues unveiled. In *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)*, pages 1199–1210. IEEE, 2019.

[4] Bilal Akil, Ying Zhou, and Uwe Röhm. On the usability of hadoop mapreduce, apache spark & apache flink for data science. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 303–310. IEEE, 2017.

[5] Morten Sieker Andreasen, Henrik Villemann Nielsen, Simon Ormholt Schrøder, and Jan Stage. What happened to remote usability testing? an empirical study of three methods. CHI '07, page 1405–1414, New York, NY, USA, 2007. Association for Computing Machinery.

[6] Apple. Apple: Differential Privacy Overview, 2023. https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf.

[7] Narges Ashena, Oana Inel, Badrie L. Persaud, and Abraham Bernstein. Casual users and rational choices within differential privacy. In *Proceedings of the 2024 IEEE Symposium on Security and Privacy*, pages 88–88, 2024.

[8] Titus Barik, Justin Smith, Kevin Lubick, Elisabeth Holmes, Jing Feng, Emerson Murphy-Hill, and Chris Parnin. Do developers read compiler error messages? In *2017 IEEE/ACM 39th International Conference on Software Engineering (ICSE)*, pages 575–585, 2017.

[9] Brett A. Becker. An effective approach to enhancing compiler error messages. In *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*, SIGCSE '16, page 126–131, New York, NY, USA, 2016. Association for Computing Machinery.

[10] Brett A Becker, Paul Denny, Raymond Pettit, Durell Bouchard, Dennis J Bouvier, Brian Harrington, Amir Kamil, Amey Karkare, Chris McDonald, Peter-Michael Osera, et al. Compiler error messages considered unhelpful: The landscape of text-based programming error message research. *Proceedings of the working group reports on innovation and technology in computer science education*, pages 177–210, 2019.

[11] Skye Berghel, Philip Bohannon, Damien Desfontaines, Charles Estes, Sam Haney, Luke Hartman, Michael Hay, Ashwin Machanavajjhala, Tom Magerlein, Gerome Miklau, et al. Tumult analytics: a robust, easy-to-use, scalable, and expressive framework for differential privacy. *arXiv preprint arXiv:2212.04133*, 2022.

[12] Nigel Bevan. Practical issues in usability measurement. *Interactions*, 13(6):42–43, 2006.

[13] John Brooke. Sus: a "quick and dirty'usability. *Usability evaluation in industry*, 189(3), 1996.

[14] Brooke Bullek, Stephanie Garboski, Darakhshan J Mir, and Evan M Peck. Towards understanding differential privacy: When do people trust randomized response technique? In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3833–3837, 2017.

[15] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 267–284, 2019.

[16] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021.

[17] Sílvia Casacuberta, Michael Shoemate, Salil Vadhan, and Connor Wagaman. Widespread underestimation of sensitivity in differentially private libraries and how to fix it. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 471–484, 2022.

[18] Lynne Cooke. Assessing concurrent think-aloud protocol as a usability test method: A technical communication approach. *IEEE Transactions on Professional Communication*, 53(3):202–215, 2010.

[19] Rachel Cummings, Gabriel Kaptchuk, and Elissa M Redmiles. " i need a better description": An investigation into user expectations for differential privacy. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 3037–3052, 2021.

[20] Damien Desfontaines. *Lowering the cost of anonymization*. PhD thesis, ETH Zurich, 2020.

[21] DiffPrivLib, 2023. https://github.com/IBM/differential-privacy-library.

[22] DP Creator, 2023. https://github.com/opendp/dpcreator.

[23] Joseph S Dumas and Janice Redish. A practical guide to usability testing, 1999.

[24] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[25] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[26] Jennifer Fereday and Eimear Muir-Cochrane. Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International journal of qualitative methods*, 5(1):80–92, 2006.

[27] Marco Gaboardi, James Honaker, Gary King, Jack Murtagh, Kobbi Nissim, Jonathan Ullman, and Salil Vadhan. Psi ({\Psi}): a private data sharing interface. *arXiv preprint arXiv:1609.04340*, 2016.

[28] Gonzalo M Garrido, Xiaoyuan Liu, Florian Matthes, and Dawn Song. Lessons learned: Surveying the practicality of differential privacy in the industry. *Proceedings on Privacy Enhancing Technologies*, 2:151–170, 2023.

[29] Google. Google: Differentially private heatmaps, 2023. https://blog.research.google/2023/04/differentially-private-heatmaps.html.

[30] Google's differential privacy libraries, 2023. https://github.com/google/differential-privacy.

[31] Douglas B Grisaffe. Questions about the ultimate question: conceptual considerations in evaluating reichheld's net promoter score (nps). *Journal of Consumer Satisfaction, Dissatisfaction and Complaining Behavior*, 20:36, 2007.

[32] Samuel Haney, Damien Desfontaines, Luke Hartman, Ruchit Shrestha, and Michael Hay. Precision-based attacks and interval refining: how to break, then fix, differential privacy on finite computers. *arXiv preprint arXiv:2207.13793*, 2022.

[33] Björn Hartmann, Daniel MacDougall, Joel Brandt, and Scott R. Klemmer. What would other programmers do: suggesting solutions to error messages. CHI '10, page 1019–1028, New York, NY, USA, 2010. Association for Computing Machinery.

[34] Bargav Jayaraman, Lingxiao Wang, Katherine Knipmeyer, Quanquan Gu, and David Evans. Revisiting membership inference under realistic assumptions. *arXiv preprint arXiv:2005.10881*, 2020.

[35] Jiankai Jin, Eleanor McMurtry, Benjamin IP Rubinstein, and Olga Ohrimenko. Are we there yet? timing and floating-point attacks on differential privacy systems. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 473–488. IEEE, 2022.

[36] Noah Johnson, Joseph P Near, Joseph M Hellerstein, and Dawn Song. Chorus: a programming framework for building scalable differential privacy mechanisms. In *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 535–551. IEEE, 2020.

[37] Daniel Kifer, Solomon Messing, Aaron Roth, Abhradeep Thakurta, and Danfeng Zhang. Guidelines for implementing and auditing differentially private systems. *arXiv preprint arXiv:2002.04049*, 2020.

[38] Patrick Kühtreiber, Viktoriya Pak, and Delphine Reinhardt. Replication: The effect of differential privacy communication on german users' comprehension and data sharing attitudes. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 117–134, 2022.

[39] Tianshi Li, Yuvraj Agarwal, and Jason I Hong. Coconut: An ide plugin for developing privacy-friendly apps. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(4):1–35, 2018.

[40] Wired Magazine. T-Mobile's $150 Million Security Plan Isn't Cutting It, 2023. https://www.wired.com/story/tmobile-data-breach-again/.

[41] Frank D McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pages 19–30, 2009.

[42] Parmita Mehta, Sven Dorkenwald, Dongfang Zhao, Tomer Kaftan, Alvin Cheung, Magdalena Balazinska, Ariel Rokem, Andrew Connolly, Jacob Vanderplas, and Yusra AlSayyad. Comparative evaluation of big-data systems on scientific image analytics workloads. *arXiv preprint arXiv:1612.02485*, 2016.

[43] Michael Meng, Stephanie Steinhardt, and Andreas Schubert. Application programming interface documentation: What do software developers want? *Journal of Technical Writing and Communication*, 48(3):295–330, 2018.

[44] Microsoft. Microsoft AI: Differential Privacy, 2023. https://www.microsoft.com/en-us/ai/ai-lab-differential-privacy.

[45] Ilya Mironov. On significance of the least significant bits for differential privacy. In *Proceedings of the 2012 ACM conference on Computer and communications security*, pages 650–661, 2012.

[46] Jack Murtagh, Kathryn Taylor, George Kellaris, and Salil Vadhan. Usable differential privacy: A case study with psi. *arXiv preprint arXiv:1809.04103*, 2018.

[47] Varvana Myllärniemi, Sari Kujala, Mikko Raatikainen, and Piia Sevoń. Development as a journey: factors supporting the adoption and use of software frameworks. *Journal of software engineering research and development*, 6:1–22, 2018.

[48] Alena Naiakshina, Anastasia Danilova, Christian Tiefenau, Marco Herzog, Sergej Dechand, and Matthew Smith. Why do developers get password storage wrong? a qualitative usability study. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 311–328, 2017.

[49] Priyanka Nanayakkara, Johes Bater, Xi He, Jessica Hullman, and Jennie Rogers. Visualizing privacy-utility trade-offs in differentially private data releases. *Proceedings on Privacy Enhancing Technologies*, 2022(2):601–618.

[50] Jakob Nielsen. *Usability engineering*. Morgan Kaufmann, 1994.

[51] Jakob Nielsen. Usability metrics: Tracking interface improvements. *IEEE software*, 13(6):1–2, 1996.

[52] OpenDP, 2023. https://github.com/opendp/opendp.

[53] Nicolas Papernot. Machine learning at scale with differential privacy in TensorFlow. In *2019 USENIX Conference on Privacy Engineering Practice and Respect (PEPR 19)*, 2019.

[54] PipelineDP, 2023. https://pipelinedp.io/.

[55] James Prather, Raymond Pettit, Kayla Holcomb McMurry, Alani Peters, John Homer, Nevan Simone, and Maxine Cohen. On novices' interaction with compiler error messages: A human factors approach. In *Proceedings of the 2017 ACM Conference on International Computing Education Research*, ICER '17, page 74–82, New York, NY, USA, 2017. Association for Computing Machinery.

[56] Associated Press. Wawa agrees to payment, security changes for '19 data breach, 2022. https://apnews.com/article/technology-pennsylvania-malware-attorney-generals-office-0ebedd8dce8bf0e21833f52944a48b56.

[57] Privacy on Beam, 2023. https://github.com/google/differential-privacy/tree/main/privacy-on-beam.

[58] Chorus Repository, 2023. https://github.com/uvm-plaid/chorus.

[59] Jayshree Sarathy, Sophia Song, Audrey Haque, Tania Schlatter, and Salil Vadhan. Don't look at the data! how differential privacy reconfigures the practices of data science. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2023.

[60] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017.

[61] Anshu Singh and Syahri Ikram. Benchmarking differential privacy python tools. https://github.com/dsaidgovsg/benchmarking-differential-privacy-tools, 2023.

[62] V Javier Traver. On compiler error messages: what they say and what they mean. *Advances in Human-Computer Interaction*, 2010:1–26, 2010.

[63] Christoph Treude, Martin P Robillard, and Barthélémy Dagenais. Extracting development tasks to navigate software documentation. *IEEE Transactions on Software Engineering*, 41(6):565–581, 2014.

[64] U.S. Census Bureau. Why the Census Bureau Chose Differential Privacy, 2023. https://www.census.gov/library/publications/2023/decennial/c2020br-03.html.

[65] Royce J Wilson, Celia Yuxin Zhang, William Lam, Damien Desfontaines, Daniel Simmons-Marengo, and Bryant Gipson. Differentially private sql with bounded user contribution. *Proceedings on Privacy Enhancing Technologies*, 2:230–250, 2020.

[66] Aiping Xiong, Tianhao Wang, Ninghui Li, and Somesh Jha. Towards effective differential privacy communication for users' data sharing decision and comprehension. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 392–410. IEEE, 2020.

[67] Aiping Xiong, Chuhao Wu, Tianhao Wang, Robert W Proctor, Jeremiah Blocki, Ninghui Li, and Somesh Jha. Using illustrations to communicate differential privacy trust models: An investigation of users' comprehension, perception, and data sharing decision. *arXiv preprint arXiv:2202.10014*, 2022.

[68] Ashkan Yousefpour, Igor Shilov, Alexandre Sablay-rolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, et al. Opacus: User-friendly differential privacy library in pytorch. *arXiv preprint arXiv:2109.12298*, 2021.

[69] ZetaSQL differential privacy extension, 2023. https://github.com/google/differential-privacy/tree/main/examples/zetasql.

A Eligibility Survey

For questions that test participants' understanding, we highlight the correct answer in bold.

Eligibility Questions after displaying IRB-approved consent form

- I have read and understood the information above. No/Yes

- I want to proceed to complete the eligibility survey for this research study. No/Yes

- Are you at least 18 years old? No/Yes

- Do you reside in the United States? No/Yes

- Have you performed statistical data analysis in Python? No/Yes

- Have you used the Jupyter Notebook before? No/Yes

- Are you willing to participate in a study to evaluate a data science tool that will require you to code in Python in a Jupyter Notebook? No/Yes

- Are you willing to participate in a 1.5-hour usability study remotely via Microsoft Teams? No/Yes

Questions on Python, DP, and basic demographics

1. How many years have you been coding in Python?

    (a) 0-1

    (b) 2-3

    (c) More than 3

2. How many years have you been using the Jupyter Notebook?

    (a) 0-1

    (b) 2-3

    (c) More than 3

3. Which of the following best describes how you use Python and the Jupyter notebook for statistical analysis?

    (a) They are my preferred language/tool

    (b) I am comfortable using them but I prefer other languages/tools (e.g., R)

    (c) I can work with them but often need to resort to documentation

    (d) I rarely use them and need additional time to get familiar with them.

4. Use "set" instead of "list" as a Python data structure for a sequence of elements when:

    (a) elements will be appended to increase the size of the sequence

    (b) the order of items is important

    (c) **it is important to know if the sequence contains a specific item**

    (d) it is important to know the item with maximum value in the sequence

    (e) I don't know

5. What is the output of the following code?

    ```
    str1 = "DataScience is fun!"
    print(str1[4:12])
    ```

    (a) **Science**

    (b) Data Sci

    (c) aScience

    (d) Error

    (e) I don't know

6. Have you heard of the term differential privacy (DP) before?

    (a) No

    (b) Yes

7. Have you ever written code to implement differential privacy (DP) in any capacity?

    (a) No

    (b) Yes

8. In differential privacy, which value of the privacy parameter ε provides stronger privacy?

    (a) **$\varepsilon = 0.1$**

(b) $\varepsilon = 1.0$

(c) I don't know

9. Releasing two differentially private statistics, one with $\varepsilon_1 = 0.1$ and the other with $\varepsilon_2 = 0.5$, results in a total privacy loss of:

(a) $\varepsilon = 0.1$

(b) $\varepsilon = 0.5$

(c) **$\varepsilon = 0.6$**

(d) $\varepsilon = 0.05$

(e) I don't know

10. If the mechanism M returns a number and satisfies differential privacy with $\varepsilon = 0.1$, does abs(M(x)) satisfy differential privacy, where abs is the absolute value function?

(a) No, not necessarily

(b) **Yes, for $\varepsilon = 0.1$**

(c) Yes, for some $\varepsilon > 0.1$

(d) I don't know

11. Which of the following is an advantage of using Differential Privacy?

(a) It guarantees complete anonymity of the data subjects

(b) It ensures that the data is completely accurate

(c) **It provides a tradeoff between privacy and utility of the data**

(d) It is a computationally simple method for preserving privacy in large datasets

(e) I don't know

12. What is your age?

13. What is your gender?

14. Are you an undergraduate or a graduate student?

B The Handout and the Tutorials

Available at Open Science Framework (OSF): https://osf.io/ag2fj/?view_only=29a9bc2a30574befa9f3d0643951b9c6

C Post-Task Survey

For questions that test participants' understanding, we highlight the correct answer in bold.

1. Please enter your participant ID

2. Please rate the following statements using the [Likert] scale indicated below.

(a) I think that I would like to use [DP tool] frequently.

(b) I found [DP tool] unnecessarily complex.

(c) I thought [DP tool] was easy to use.

(d) I think that I would need the support of a technical person to be able to use [DP tool].

(e) I found the various functions in [DP tool] were well integrated.

(f) I thought there was too much inconsistency in [DP tool].

(g) I would imagine that most people would learn to use [DP tool] very quickly.

(h) I found [DP tool] very cumbersome to use.

(i) I felt very confident using [DP tool].

(j) I needed to learn a lot of things before I could get going with [DP tool].

(k) I found [DP tool] introduced DP concepts appropriately for me to perform the tasks.

(l) I feel I have to learn DP concepts more systematically to solve the tasks.

3. If another data scientist that you know needs to use differential privacy in their data analysis, how likely is it that you would recommend Tumult Analytics to them?

• 10-point Likert scale, "Not at all likely" to "Extremely likely"

4. If you completed at least one task in the study, what helped you successfully complete the task(s)? Choose all that apply.

(a) The Differential Privacy handout (including the video)

(b) The [DP tool] tutorial (including its examples)

(c) The official [DP tool] documentation

(d) My prior data science skills (like Python, Pandas, statistics, etc.)

(e) My prior knowledge of Differential Privacy

(f) Other (please specify)

(g) N/A (I didn't complete any tasks)

5. What hindered your completion of the tasks? Choose all that apply.

(a) The Differential Privacy handout (including the video)

(b) The [DP tool] tutorial (including its examples)

(c) The official [DP tool] documentation

(d) My prior data science skills (like Python, Pandas, statistics, etc.)

(e) My prior knowledge of Differential Privacy

(f) Other (please specify)

6. If the mechanism M returns a number and satisfies differential privacy with $\varepsilon = 0.1$, does abs(M(x)) satisfy differential privacy, where abs is the absolute value function?

(a) No, not necessarily

(b) **Yes, for $\varepsilon = 0.1$**

(c) Yes, for some $\varepsilon > 0.1$

(d) I don't know

7. In differential privacy, which value of the privacy parameter ε provides stronger privacy?

(a) **$\varepsilon = 0.1$**

(b) $\varepsilon = 1.0$

(c) I don't know

8. Releasing two differentially private statistics, one with $\varepsilon_1 = 0.1$ and the other with $\varepsilon_2 = 0.5$, results in a total privacy loss of:

(a) $\varepsilon = 0.1$

(b) $\varepsilon = 0.5$

(c) **$\varepsilon = 0.6$**

(d) $\varepsilon = 0.05$

(e) I don't know

9. Which of the following is an advantage of using Differential Privacy?

(a) It guarantees complete anonymity of the data subjects

(b) It ensures that the data is completely accurate

(c) **It provides a tradeoff between privacy and utility of the data**

(d) It is a computationally simple method for preserving privacy in large datasets

(e) I don't know

D Post-Task Interview

For questions that test participants' understanding, we give sample correct answers after each question.

Thank you for completing/making an effort to complete the tasks with [tool] and the post-task survey. Now we have a few questions for you to reflect on your experience with the study.

1. After completing the tasks, can you explain differential privacy to me in your own words?
Correct answer: Differential privacy is a formal property that limits the distributional difference between a statistic computed on one dataset and the same statistic computed on a neighboring dataset.

2. Consider the tasks you worked on just now, can you explain:

(a) What was the privacy budget for each task?
Correct answer: Depends on the parameters used by the participant—it should be equal to the value of ε used in each task's solution.

(b) What was Epsilon?
Correct answer: Same as (a).

(c) What was the total privacy budget for the whole notebook?
Correct answer: $3 \times$ (answer from (a)), by sequential composition.

(d) If the results you computed were released to the public, how strong would you expect the privacy protection for individuals in the original data to be?

3. During this study, what helped you most in understanding the concepts (e.g., privacy budget, Epsilon) that we discussed just now? Please rank the following options from "most useful" to "least useful".

(a) The Differential Privacy handout (including the video)

(b) The [DP tool] tutorial (including its examples)

(c) The official [DP tool] documentation

(d) My prior knowledge of Differential Privacy

(e) Other (please specify)

4. When using [tool] in the study, what aspects/components of [tool] do you think are helpful for you to complete the tasks?

5. When using [tool] in the study, what aspects/components of [tool] do you think are frustrating for you to complete the tasks?

6. After this study, what recommendation(s) do you have to improve the usability of [tool]?

7. Can you tell us what helped you successfully complete the task(s)?

(a) The Differential Privacy handout (including the video)

(b) The [DP tool] tutorial (including its examples)

(c) The official [DP tool] documentation

    (d) My prior data science skills (like Python, Pandas, statistics, etc.)

    (e) My prior knowledge of Differential Privacy

    (f) Other (please specify)

8. Can you tell us what hindered your completion of the (task)?

    (a) The Differential Privacy handout (including the video)

    (b) The [DP tool] tutorial (including its examples)

    (c) The official [DP tool] documentation

    (d) My prior data science skills (like Python, Pandas, statistics, etc.)

    (e) My prior knowledge of Differential Privacy

    (f) Other (please specify)

E   Task Dataset

The dataset used for the tasks was provided to participants in a CSV file. This is a synthetic dataset that counted restaurant visits across a week, where each record represented a distinct visit with a visitor ID. The full dataset is available at Open Science Framework (OSF): https://osf.io/ag2fj/?view_only=29a9bc2a30574befa9f3d0643951b9c6

F   Task Solutions

We wrote sample solutions for the three tasks from Table 2 for each tool we studied. Participant solutions were usually similar, but not necessarily identical. We will make these sample solutions available publicly via Open Science Framework on publication of the paper.

G   Additional Figures
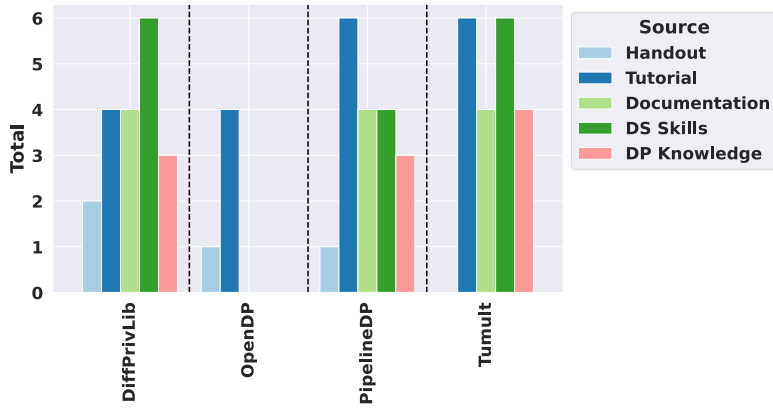
(a) By Tool
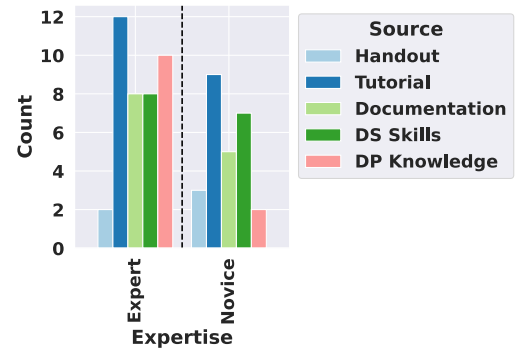
(b) By Expertise

Figure 5: Factors helping task completion by tool and expertise.
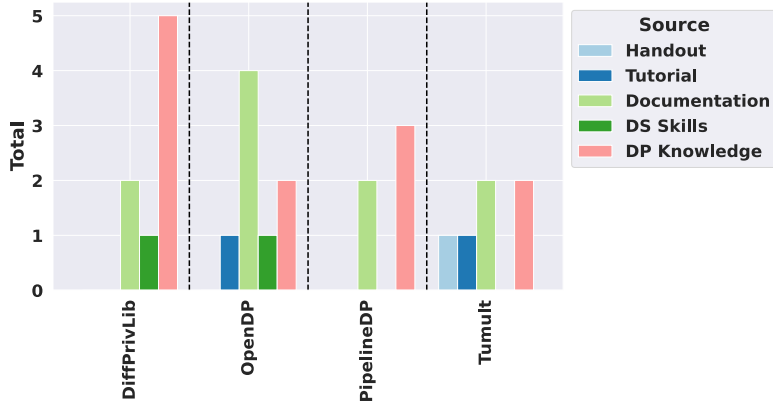
(a) By Tool
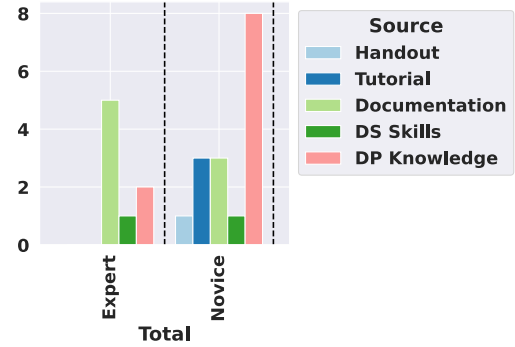
(b) By Expertise

Figure 6: Factors hindering task completion by tool and expertise.
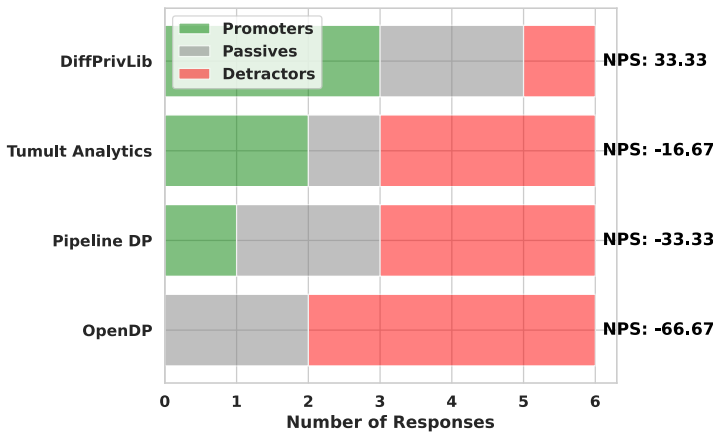
(a) Net Promoter Score (NPS) results
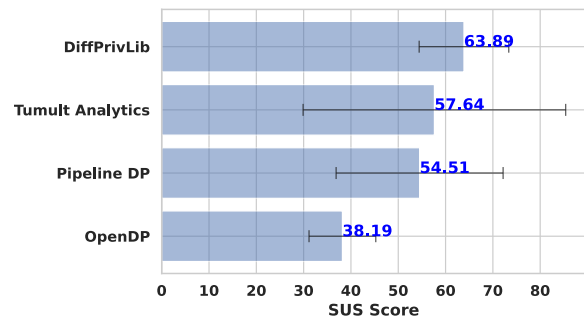
(b) System Usability Scale (SUS) results

Figure 7: User satisfaction scores: (a) Net Promoter Score (NPS), and (b) System Usability Score (SUS).

Navigating Autonomy: Unveiling Security Experts' Perspectives on Augmented Intelligence in Cybersecurity

Neele Roch
ETH Zurich

Hannah Sievers
ETH Zurich

Lorin Schöni
ETH Zurich

Verena Zimmermann
ETH Zurich

Abstract

The rapidly evolving cybersecurity threat landscape and shortage of skilled professionals are amplifying the need for technical support. AI tools offer great opportunities to support security experts by augmenting their intelligence and allowing them to focus on their unique human skills and expertise. For the successful design of AI tools and expert-AI interfaces, however, it is essential to understand the specialised security-critical context and the experts' requirements. To this end, 27 in-depth interviews with security experts, mostly in high-level managerial roles, were conducted and analysed using a grounded theory approach. The interviews showed that experts assigned tasks to AI, humans, or the human-AI team according to the skills they attributed to them. However, deciding how autonomously an AI tool should be able to perform tasks is a challenge that requires experts to weigh up factors such as trust, type of task, benefits, and risks. The resulting decision framework enhances understanding of the interplay between trust in AI, especially influenced by its transparency, and different levels of autonomy. As these factors affect the adoption of AI and the success of expert-AI collaboration in cybersecurity, it is important to further investigate them in the context of experts' AI-related decision-making processes.

1 Introduction

The growing dependence on digital devices, services, and data for daily tasks by individuals, companies, and governments increases productivity but concurrently increases the vulnerability to cyberattacks. Cybercrime is growing expo-

nentially, with organisations experiencing an average of 1248 attacks per week [10, 11]. This results in a fast-paced and demanding work environment for cybersecurity teams. Simultaneously, organisations face a significant shortage of cybersecurity experts (CSEs) to cope with increasing security-related demands. In 2022 the estimated cybersecurity workforce gap stood at 3.4 million jobs globally [42], and existing CSEs report high work stress levels, fear of burnout, and feeling set up for failure in a chronic state of work overload [4, 41]. While educating future CSEs is an essential task, it may take a long time and currently does not catch up with the increasing demands.

Hence, technical solutions, such as Artificial Intelligence (AI) are a promising approach to ensure the secure operability of IT systems, to compensate for the current lack of CSEs, and relieve experts. AI tools have the potential to support human CSEs by enhancing and strengthening the human's abilities to act, analyse, decide, see and hear [61]. Human-AI collaboration does not aim to replace humans, but "to achieve complex goals by combining human and AI, thereby reaching superior results to those each of them could have accomplished separately [...]." [25, p. 640]. This complementary collaboration has already been proven to be successful in other domains, such as for medical [13, 14] or military use cases [23]. Hence, the exploration of this unique collaboration could provide similar benefits in the high-stakes and complex work field of cybersecurity.

Due to its technical disposition, cybersecurity is a field where AI could be applied to a range of use cases; especially in the domain of detection and response, e.g., for continuous security monitoring [29, 49, 56, 68, 72, 101]. Other cybersecurity research is concerned with, e.g., the technical development of automated calculation of risk scores [76, 91], inferring the probability for a security incident [71], or dark web investigations for threat intelligence and text analysis [5, 28, 43].

However, not all cybersecurity tasks and decisions, such as assessments or stakeholder communication, can be easily automated. For example, tasks such as risk assessments require

an understanding of the organisation's strategy and careful consideration of and communication with all stakeholders [44]. Hence, understanding the context, the experts' needs, and requirements for the intended collaboration is essential as it is influenced by the specific tasks, the required skills, and desired behaviours [79]. Furthermore, human-AI collaboration has often only been studied through the lens of novice users and non-specialised tasks; yet, experts have different requirements for working together with AI systems, as they, e.g., are more knowledgeable or self-confident in their domain of expertise and more averse to algorithmic advice [8, 12, 36, 70].

To provide a better understanding of the potential for expert-AI collaboration in cybersecurity, and the CSEs' requirements and perceptions, we conducted interviews with *N*=27 security professionals in Chief Information Security Officer (CISO) or related roles. As this research was exploratory with a focus on gaining an in-depth understanding, a qualitative interview approach using grounded theory was chosen.

The first objective of this research was to get a good understanding of security experts' tasks, and responsibilities, and to explore the potential CSEs see in collaborating with AI to complete their tasks, leading to the first research question (RQ).

RQ1: *What tasks and responsibilities do cybersecurity experts have, and which can be augmented by AI tools?*

We found that the experts' responsibilities are concerned with designing, implementing, and constantly reviewing and improving their organisations' cyber and information security strategies. The experts confirmed that, based on their assessments of the task nature, and both team members' capabilities, various tasks could be automated or augmented by AI.

After identifying potentially suitable tasks for collaboration, it was essential to understand the factors that are relevant for the successful collaboration of experts with AI in cybersecurity, e.g., the security experts' specific strengths, their requirements, and willingness related to the collaboration, and the capabilities they attribute to AI.

RQ2: *What is the cybersecurity experts' perspective on collaborating with AI tools to complete their tasks?*

The experts believed that AI can improve their workflows and relieve them of extensive and repetitive tasks, but also help with more discretionary tasks. Tasks such as monitoring, and analysing big amounts of data, or alerting the experts in unusual cases were identified as use cases where experts can benefit from the use of AI. On the contrary, especially stakeholder communication and the integration of the organisational context into the final decisions should remain in the human experts' domain.

Finally, user perceptions such as trust towards and concerns related to using AI tools have been shown to be relevant for the tool's acceptance and successful collaboration [18]. Of

particular focus is how these might change across varying levels of autonomy and automation of AI tools concerning the level of transparency, autonomy, and adaptability of a system. AI tools can range from decision support tools, over collaborative scenarios where human approval is required, and situations where the human can only veto, to fully automated task execution by the AI.

RQ3: *What are the cybersecurity experts' perceptions on automation, autonomy, and trust in expert-AI collaboration?*

We found that deciding how autonomously an AI should be able to act was a difficult decision, which required experts to weigh factors such as their trust in AI tools and the suitability of the task for AI or the human expert, which influenced their assessment of deploying AI at different autonomy levels.

This research yielded valuable insights into the collaboration potential of AI and CSEs, which can guide the design of AI tools in cybersecurity. These tools can help fill the workforce gap by augmenting the existing CSEs' with AI, to free them up for other tasks that require their unique human capabilities and expertise. To that end, we provide two contributions:

First, within the high-stakes environment of cybersecurity, we explored for which type of tasks CSEs seek support from AI and outline how experts decided to share tasks between themselves and AI tools.

Second, we present a comprehensive autonomy decision framework that describes how the interplay of factors like the nature of the task, trust in AI, and a risk-benefit assessment impacts the decision to utilize AI on different autonomy levels. It provides a structured approach to determining the appropriate balance between human expertise and AI autonomy in cybersecurity.

These contributions help advance the understanding of expert-AI collaboration in cybersecurity and can guide the practical implementation of collaborative interfaces for CSEs and AI, fostering more effective and secure cyber defence strategies.

2 Related Work

In the following, we present related work concerned with AI in cybersecurity and expert-AI collaboration.

2.1 AI in Cybersecurity

Currently, the use of AI cybersecurity tools is relatively rare, with only one-third of organizations using or planning to use them [92]. Among these, 80% use AI to detect, 64% to predict, and 55% to reactively mitigate cyberattacks [92]. The primary applications for AI in cybersecurity are network security (75%), data security (71%), and endpoint security (68%) [15]. Diverse AI methods like machine learning (ML), or

natural language processing (NLP) are promising for application in cybersecurity [77]. However, the use of AI also introduces novel risks, e.g., when ML models are tampered with during training through poisoning attacks, manipulating AI predictions [89]; or hallucinations, where LLMs produce syntactically correct answers that are nevertheless made up and false [6, 78].

Kaur et al. [46] identified literature concerned with the integration of AI into various use cases in cybersecurity and classified them according to the five NIST cybersecurity framework functions: identify, protect, detect, respond, and recover[1].

For the function of *identifying*, there are several studies concerned with supporting individuals and organisations to identify threats and risks, e.g., through automated calculation of risk scores [76, 91], or by inferring the probability for a security incident with AI [71]. The availability of related tools in practice is comparably mature: Tenable's Exposure AI [87], an attack surface management tool, or IBM's Guardium [39], which offers functionalities for risk assessment, vulnerability scanning and patch management. In the function of *protection*, AI can be used for threat simulation to identify and address gaps in software or misconfigurations in settings. When using AI in this function, e.g., for threat intelligence, it can efficiently combine data from multiple sources such as networks, users and IoT devices, for real-time monitoring and analysis [64]. Industry solutions supporting this function are IBM's Verify [40], offering AI functionalities for managing digital identities and access rights, and Zscalers Data Protection [102], providing data classification and visibility for the locations of sensitive data using AI. For the *detection* of anomalies, AI can increase event detection rates and detect unknown threats, as is demonstrated in AI tools for continuous security monitoring [29, 49, 56, 68, 72, 101]. Due to AI tools being able to monitor significantly more data, incidents can be reported more quickly and effectively [83]. Tools such as Microsoft's Security Copilot [62] support threat detection and prevention, and Tessian's Complete Cloud Security Email platform [88] uses AI to detect phishing and protect sensitive data on email. Once a cybersecurity incident is detected, AI tools can be used for the *response*, e.g., for automated isolation of affected entities [30, 74], or the automated remediation, such as preventing the spread of malware [38] or recommendations for countermeasures [67]. Industry solutions for the response function include Darktrace [22], which offers an automated response to threats such as ransomware, or Malwarebytes [55] and Kaspersky's Endpoint Security [45], offering AI-based malware identification and detection. The use of AI in the function of *recovery* is less mature, and the amount of research in this area is still relatively small compared to the other functions. Nevertheless, using AI for the aggregation of incidents [16], or the concluding analysis of vulnerabilities [60] are two examples of solutions

in this function. Industry solutions for the recovery after a successful cyberattack are comparably scarce, however, Darktrace [22] offers an AI-based functionality that supports tasks in this domain, such as incident reporting.

Kaur et al. [46] emphasized the impact of human-AI collaboration in developing practical and usable AI for cybersecurity. Despite existing literature and AI tools from practice, the alignment with CSEs' needs, acceptance, or trust is rarely evaluated. Our research addresses this gap by examining experts' perspectives on AI adoption in cybersecurity, and identifying potential matches and mismatches with current tools and proposals.

2.2 Expert-AI Collaboration

Effective collaboration between humans and AI has been shown in research [25] and yet, the success of this collaboration depends on the understanding of delegation dynamics [8, 36, 70], is influenced by attitudes towards and knowledge of AI [70], and further factors such as overcoming cognitive biases [73] or algorithmic aversion which domain experts are especially prone to [12, 63, 100]. Supporting human experts with AI capabilities is based on the assumption that both actors can bring complementary skills to the collaboration, enhancing overall performance. While humans are suited for social tasks and unexpected situations, AI can perform repetitive and monotonous tasks quickly, accurately and reliably [35, 50]. In particular, AI can collect information logically and arithmetically and then process large amounts of data by weighting, prioritising, analysing and combining it [50]. In contrast, human actors can rely on their senses, emotional intelligence and social skills to build relationships and motivate employees [9, 23, 35]. Unlike AI, humans are able to use their intuition, creativity, and common sense in situations they were not trained for [35], enabling them to creatively develop solutions even in open and unfamiliar situations [48, 50, 94]. However, while existing research of human-AI collaboration yields insightful findings, they often stem from generalizable tasks performed by novices and might not hold in high-stakes environments or those that require expert knowledge as the domain of cybersecurity does [8, 36, 70]. Initial insights in the teaming of experts and AI show that especially in data-heavy fields, like medicine or military, where decisions are discretionary, rather than ruled by clear guidelines, expert-AI collaboration is beneficial to performance [2, 13, 24, 52]. While experts excel in unstructured environments, too many decision variables overwhelm human processing [3]. AI can reduce this complexity, enabling human experts to make informed decisions. Research on behaviours, skills, and abilities required for successful human-AI collaboration is fragmented and task-dependent [79], motivating our research to look specifically at the cybersecurity domain. As the context, the task, and the target group influence human-AI collaboration [79] we try to understand how CSEs and AI can collaborate across different

---

[1] https://www.nist.gov/cyberframework/getting-started/online-learning/five-functions

tasks. To do so, we introduce related frameworks describing relevant aspects of human-AI or related human-automation interaction in the following section.

2.3 Human-AI Collaboration Frameworks

Human-AI frameworks compare and describe interactions across various dimensions, such as trust [93], autonomy [37], or the influence of psychological factors [85]. For this purpose, Salikutluk et al. [75] evaluated the desired autonomy of a physical AI agent in a shared workspace, based on the situational context and the human agent's self-confidence, effects of task failure, understanding human capabilities (i.e., theory of mind), comparison of AI and human competence, and whether the human agent needs to adjust their actions. After refining the factors in a pre-study, they established that participants prefer an adapted autonomy level based on those factors, compared to fixed autonomy levels.

Similarly, Simmler and Frischknecht [80] derived a taxonomy describing human-AI collaboration that relies on two gradual parameters: *autonomy*, capturing the intractability of the system's actions, and *automation*, capturing the human level of control over the system's actions. The taxonomy defines levels of automation, reflecting the extent of human involvement in task execution and how autonomously an AI can act. At the lowest level, AI functions as a decision support tool, relieving experts of preparatory work for decision-making. Contrary, on the highest level five, there is no interaction between the human and the AI, with the AI being fully independent. Table 3 in Appendix A describes the five levels in more detail. The four key characteristics that define an autonomous system according to [80] include the system's transparency, determinism, adaptability, and openness (see Table 2 in Appendix A for a description). Overall, the taxonomy can be used to assess and describe the roles and responsibilities of each actor in human-AI collaboration, and therefore influenced the design of the AI autonomy section in our interviews.

The importance of trust for humans working with autonomous agents and the technology's acceptance has been shown in other research and motivated the trust section of our interviews [59, 81, 82, 97]. Research has shown that humans interact with machines differently than with other humans, nevertheless, there are similarities. For instance, humans apply human social rules and behaviour to machines [33, 34, 66], and trustor- and context-related factors have been considered equivalent regardless if the trustee is human or a machine [34]. In the organisational context, trust in technology describes the users' belief in the system's ability to perform as expected and the users' willingness to depend on the technology and make themselves vulnerable in uncertain and risky situations [31, 57, 65]. Theoretical descriptions of the human-technology trust relationship assume trust evolves over time; and, begins with an initial trust level even prior to the first

interaction [65, 86]. This is then either confirmed or refuted upon the initial interaction with the technology [65, 86]. Succeeding the initial interaction, the users' trust was found to be positively related to the system's usability, and during the early periods also to the system's reliability [65]. When users do not understand the AI's prediction, and it conflicts with the implications of their own mental model, it can lead to uncertainties regarding decision-making [47]. One approach to support trust between humans and AI is the facilitation of transparency [32, 69, 97]. The provision of the AI's reasoning has shown to have a positive effect on cognition-based trust, indicating that this transparency can support bridging the discrepancies between the human mental model and the AI model and fostering the collaborative performance [97]. Therefore, through our interviews, we explored how experts would build trust in AI systems, and which factors could strengthen and weaken their trust.

Overall, the interplay of experts and AI needs to be better understood to enhance their collaboration, the expert's ability to evaluate the AI and its predictions, and the communication between the AI and the expert. Our study contributes to this understanding by exploring the CSEs' perspective on AI adoption in cybersecurity through in-depth interviews with CISOs and related job roles that additionally make use of existing frameworks such as the autonomy-automation taxonomy [80] to put the insights into a meaningful context.

3 Method

The following section describes the interview procedure, the data analysis procedure following a grounded theory approach, the sample and ethical considerations.

3.1 Interview Procedure

Through an online survey sent prior to the scheduled interview, the experts were informed about the data collection, processing, and storage and asked to give their consent to the explained procedures. To accompany the insights of the interviews, the survey contained a general attitude towards AI scale based on the *General Attitudes Towards Robots Scale (GAToRS)* [51], adapted to the AI-specific case and provided in Appendix C. Additionally, we asked experts to fill out the *Human-Computer Trust (HCT) scale* [54] in the pre-interview survey. The interviews were mainly conducted through Zoom or Microsoft Teams, with some in-person interviews. The interviews were audio-recorded after again obtaining verbal consent from the expert. The audio files were auto-transcribed with Trint [90], and then validated manually by two researchers. Data collection, transcription, and analysis occurred concurrently, in line with the grounded theory approach [21, 84].

The semi-structured interviews consisted of three sections. The full interview guide can be found in Appendix D.

1. The first section concerned understanding the CSEs' tasks, responsibilities, and the need for support.

2. The second section explored which type or level of expert-AI collaboration would be suitable for cybersecurity tasks, including a reflection on human and AI capabilities. This section was accompanied by an explanation of the automation levels [80] (detailed in Table 3). Experts went through their tasks and were asked to elaborate on the level of automation that was suitable for the integration of AI for each task. Then, CSEs were asked to sort the tasks into a matrix considering the feasibility and desirability of AI integration, similar to a How-Now-Wow matrix[2] as illustrated in Figure 2 in Appendix D.

3. The last part of the interview focused on understanding the CSEs' trust in and perceptions of AI tools, including their hopes and worries regarding expert-AI collaboration.

Interviews were internally piloted twice, to ensure the clarity of the questions, the feasibility of the methods and to approximate the interview duration. After piloting, the interview guideline was revised to reduce the number of questions, to avoid redundancy and, to change the order of questions in the section on AI and human capabilities. The final version can be found in Appendix D.

3.2 Data analysis

The survey data was analysed descriptively to contextualize the topic and obtain a comprehensive impression of CSEs' attitudes and trust towards AI.

To analyse the interview data, we followed a grounded theory approach [21, 84], which is suitable when little is known about the topic and allows synthesizing qualitative interview data and generating research assumptions and frameworks [19]. The interview guideline aimed to elicit diverse responses on the participants' perceptions initially, but provided guidance to experts, prompting them to evaluate AI in the context of specific cybersecurity tasks. Therefore, responses focused on similar tasks and allowed the emergence of consistent patterns.

We used the central element of grounded theory, ongoing memoing, in the transcription and during all coding phases, to capture impressions and ideas. Memoing describes the process of recording thoughts, analytical insights, decisions, and ideas in relation to the research process [21]. During the coding phase, we added the technique of diagramming [21], i.e., creating visual representations of interrelations between codes, to support the development of the categories and their relationships and interactions. The coding process was structured as follows: the initial open coding phase aimed at initial codebook development, where the first five interviews were coded with a line-by-line approach. Once an understanding of underlying themes in the data was developed, line-by-line codes were transformed into incidents.

The intermediate coding process focused on axial coding. Strauss and Corbin defined axial coding as *"a set of procedures whereby data are put back together in new ways after open coding by making connections between [and within] categories"* [21] (p.96). We captured the relationships of the arising themes and their contexts by diagramming and generating situational maps [20]. During axial coding, two researchers went through the interviews and developed a situational map depicting the codes until no new codes or relationships could be added.

In the final coding phase, the codebook derived from the situational maps was transferred to the coding software MAXQDA (v24.1.0) [95]. The interviews were then coded topic by topic, and interview sections discussing the same topics were compared between participants to enrich themes and provide different variations of one topic and respective codes. The full codebook can be found as an online appendix on https://doi.org/10.3929/ethz-b-000674517, and additional details are given in Appendix B.

3.3 Sample & Recruiting

The CSE sample was mainly recruited through purposive sampling to ensure that the participants matched the target group in terms of roles and level of expertise. Experts required a minimum of 2 years of industry experience in cybersecurity roles. As peer identification is another way to determine expertise, we used additional snowball sampling to help us recruit further relevant candidates for our interviews through interviewees and recruited one out of all 27 experts through this method. Experts were approached through social media platforms, specific expert forums, mailing lists, and peers. A total of 27 security experts from 23 different organisations were interviewed between November 2023 and January 2024, and their demographics are summarized in Table 1. The experts on average had M=15.37 years (SD=8.08) of experience in the field of cybersecurity. Throughout the rest of this paper, experts will be referred to as *ME*, when they are in a managerial role; *OE*, when they are in an operational role; and *CE*, when they are in a consultancy role. The individual years of experience for the experts can be found in Table 6 and the sample scores of the AI-adapted GAToR scale [51] are shown in Table 5 in Appendix C. Overall, on a 7-point Likert scale experts tended to agree with having personal experience with AI (M=5.50, SD=1.10), and being familiar with AI (M=5.54, SD=0.65). Additional information on the experts' organisations can be found in Appendix E.

---

[2] https://gamestorming.com/how-now-wow-matrix/

Gender		Role	
Male	25	Chief Information SO[3]	16
Female	2	Information SO[3]	2
Age		Chief SO[3]	2
25-34	4	Head of Security	2
35-44	7	Junior Information SO[3]	1
45-55	13	Security Architect	1
55-64	3	Network Security	1
HCT		Security Consulting Engineer	1
M=3.71 $(SD$=1.24$)$		Manager	1

Table 1: Expert demographics, $n = 27$

3.4 Ethical Considerations

Our institution's ethics board approved the study design, following established ethical guidelines for psychological research involving humans [7]. By collecting age ranges instead of a concrete age, we minimized the potential for privacy invasion. An informed consent form explaining the purpose of the study, data collection, and processing was given to participants before the interview. Participants were free to refuse participation and request the deletion of their data at any time without negative consequences. The audio data was transcribed, anonymized, and then deleted. The participants were given equal compensation and had the option of taking or donating the money to a charity of their choice.

4 Results

This section first describes the experts' responsibilities and tasks before summarizing the main themes that emerged in the interviews regarding the use of AI in cybersecurity, preferences for expert-AI task division, and trust in AI.

4.1 Cybersecurity Expert Roles, Responsibilities and Tasks

The experts described that their main responsibilities lie in the tactical and technical management of an organisation's information security, including the strategic information security orientation and ensuring adherence to regulatory and compliance requirements. The organisation's specific cyber and information security strategy is set out in policies, guidelines, and frameworks that are regularly reviewed, improved and audited. Experts mentioned responsibilities spanning from raising security awareness in the organisation to risk management, where threats to the organisation's assets are identified, evaluated, assessed, and appropriately treated. To protect the organisation from cyberattacks, experts described being responsible for technically steering the incident and vulnera-

---
[3]SO = Security Officer

bility management, which includes planning, designing, and implementing technical systems, such as security information and event management (SIEM). Specifically, experts described communicating cybersecurity issues to management and other employees in an appropriate and sensitive manner, briefing them in the event of a crisis, and being available to answer ad-hoc questions. To ensure information security, experts also guide, plan and support the implementation of the organisation's security infrastructure. Expert responsibilities depend on their roles, where operational positions verify incidents, notify responsible parties, and facilitate device and entity recovery, bridging technical and human aspects within organisations. Experts in in-house managerial roles are primarily employed to fulfil the tasks described above, however inevitably get involved into more operational tasks if necessary. Experts in consultancy positions advise and consult external clients, and are less involved in the implementation and focus on the strategic and governance aspects, but held expertise in similar domains without hands-on involvement. Overall, the experts described the integration of AI into cybersecurity tasks as desirable, however, the tasks' suitability for AI integration varies based on AI and human capabilities.

Experts should plan, strategize and grasp the context. Tasks that participants considered lying in their responsibility often included competencies related to **strategy, planning, and assessment**. Experts considered these tasks to be more appropriate for humans, as "*transferring that to your context and to your company and to the risk that you have in your company. That's something that [AI] simply can't do that well.*" (CE24). Additionally, if there were no formal criteria that could be embedded into an AI to determine appropriate decisions or actions for specific scenarios, experts believed that humans should be responsible. Since those discretionary decisions would be "*something that has to do with emotion. It has to do with experience, it has to do with [...] circumstances and environmental influences. That's something that humans naturally take into account that a machine can't do.*"(ME5). For difficult discretionary decisions, with incomplete information, experts mentioned the need to weigh up equally valid options and in many cases relied not only on objective factors but also on experience from unrelated areas and their gut feeling. Typically, taking responsibility "*whether the activity, which can then be very drastic, makes sense in this context, this should still be a human who ultimately bears the responsibility for what happens.*" (ME8). While the integration of AI into these tasks was rated as somewhat desirable, most experts argued that this was not feasible at this time.

AI, the expert's little helper. However, this did not imply that the use of AI for these responsibilities was always deemed out of place. Many experts recognized the advantages of using AI to **prepare** tasks, such as assessments, audits, or guiding the development of strategies. Experts felt it was

suitable for AI to **gather information** on specific topics, **summarise** existing documents, and generate preliminary reports to guide the experts' decision-making. They recognised that AI could make vast amounts of information and data accessible to them through summaries. Leveraging AI's ability to bring "*information together and combine it in such a way that it makes some sense*" (ME20) to support decision-making during assessments was deemed to add value; "*[a]lthough a verification is then necessary.*" (ME20). Cybersecurity tasks in this category were gathering information about vulnerabilities, and incidents, giving insights into the current legal landscape or preparing information for risk assessments.

The age of generative AI.   Experts also found the integration of AI useful for **generative** tasks. Most experts felt that letting AI write at least drafts for policies, frameworks, and guidelines would greatly support their work, but should be closely monitored and verified. They believed that AI would quickly learn the structure of such documents and could easily and efficiently reproduce them given different parameters to fit the respective organisation's needs. Additionally, experts mentioned that this would "*help me to create a proper language*" (ME7) for such compliance documents. The recent popularity and success of LLMs made them optimistic that AI "*will probably deliver a great policy*" (ME2), but also raised concerns that "*[AI] doesn't know my company, so it doesn't know the needs, demands, and requirements of my company*" (ME2). The experts were obviously aware of AI's fallibility, especially in regard to hallucinations, and insisted that no AI-generated document should just blindly be used but always be examined for context, and validity.

Can AI do awareness measures?   Experts could not agree on how feasible or desirable the use of AI would be to raise cybersecurity **awareness**. Some experts believed that interpersonal interactions play a major role in the area of awareness, which makes the integration of AI counterproductive. Other experts elaborated on how they could use AI for raising awareness, "*for example, [to] produce an awareness training program tailored exactly to our needs: This is how I store data, make a short video about it, there's a quiz to review what has been learned*" (ME25). Some experts further mentioned that they were already using AI for awareness measures.
In summary, the human experts should be the ones driving the process, planning an awareness strategy and evaluating the achievement of their goals. AI could be utilized *within* the awareness measures, e.g., for generating and sending phishing emails to train employees, or generating content, such as texts, graphics, and videos for training purposes.

To communicate through AI or not to communicate through AI.   Another task group that experts were reluctant to delegate was **communication**. While "*[AI] can make a*

lot of tasks easier, such as answering my emails, which I no longer want to do myself, and I am convinced that I could easily hand over to AI" (ME27), other kinds of communication required the human experts' unique qualities. Communication, especially "*when it comes to crisis communication and so on, I would also say that you also need a good deal of empathy and political skill*" (ME23). Understanding and addressing the other party's needs through communication relies heavily on interpersonal aspects and sensitivity, which are rarely explicitly visible or graspable. Therefore, tasks such as crisis management and communication with customers or legal representatives were still considered best-suited for the human expert. The integration of AI was found to be undesirable and even infeasible.
However, answering employees' security-related questions was very desirable and feasible to delegate to AI. They pictured a compliance expert chatbot which would have access to their organisation's policies and implementations and thus be able to answer ad-hoc questions compliantly and tailored to the enquirer's role and level of expertise.

Leveraging AI capabilities to protect cybersecurity.   The use of AI for **protection and prevention** was considered to make the experts' work easier and provide valuable support. While tasks like "*pen-testing requires a certain level of human intelligence, which is why it is so expensive and takes longer*" (ME20) , they expressed "*[i]f it were feasible, of course, [...] then it would be nice if it were as fully automated as possible. And then at the end you have a report [...]*" (ME20). In addition, AI could automatically review policy implementations or measures and identify potential gaps - "*tackle configuration management, verify firewall rules, check technical configuration elements*" (ME3) and potentially prevent successful attacks by "*enforcing requirements, [and] checking them*" (ME13).

The data-intensive and repetitive territory of AI.   Tasks in the domain of **detection and response** were often mentioned as desirable and feasible for AI delegation. In particular, to "*analyse large volumes of data in a very short time, while incorporating historical data*" (ME4). In many cases, experts were aware that AI, compared to humans, can analyse data more precisely and differentiate benign and malignant patterns. In addition, experts noted that AI is not affected by human disadvantages: AI has no biorhythm and can therefore perform "*24/7/365 and at an alarming speed*" (ME18). Its impartiality and absence of emotions ensure consistent results and objectivity. Moreover, many experts were aware that an AI can "*reduce the sheer volume of data*" (ME8) and constantly "*learns more and more over time*" (ME2). The use of AI for monitoring could therefore replace and facilitate the work of many human analysts.
Although the experts generally viewed the use of AI for technical monitoring positively, monitoring humans raised

ethical and legal concerns, making it undesirable. Letting the "*AI go there fully automatically and address or correct any misbehavior*" (ME2) was dismissed by experts.

The automated "*analysis for anomalies, performing baselin[ing] and outlier detection*" (ME9) was also found to be one of the most desired tasks for the integration of AI. However, the automatic response to detected anomalies, e.g., responding to suspected attacks by isolating entities, or automatically fixing detected vulnerabilities, involved more complex judgements. While the desirability of automatic responses was high, many experts were unsure about how much autonomy AI should have in these cases.

4.2 AI Autonomy in Cybersecurity

Experts were asked to elaborate on task division between them and AI, as well as the autonomy of AI in relation to the experts' previously described tasks. It is important to note that the concepts of AI automation and AI acting autonomously were used interchangeably during the experts' descriptions. As autonomy appeared to encompass aspects of automation, we will refer to the concept as "AI autonomy" from here on and detail the factors that experts deemed relevant for deciding on AI autonomy. The definition of autonomy as the "*quality or state of being self-governing*" [58] closely maps the experts' descriptions. Several experts were able to formally describe their internal thought process when evaluating AI autonomy levels for different tasks, as visualized in Figure 1. Deciding how autonomously AI should be able to act was a complex consideration for the experts. Factors such as the characteristics of the task, trust in AI, and risks and benefits played a role in this process. In the following sections, we first describe what factors experts evaluated regarding the task and their trust in AI. As shown in Figure 1, these two factors directly influence the experts' perceptions of the risk and benefit of different levels of AI autonomy. We will then describe the experts' assessments for the risks and benefits that influence what level of AI autonomy is deemed appropriate by CSEs (see Figure 1). This section concludes with the considerations for expert-AI task division at different levels of automation.

Considering the capability-task fit and the urgency. The nature of the task, and in particular how suitable experts **assessed** their **human capabilities** or **AI capabilities**, played into the decision for the AI autonomy level. For example, analytical tasks that require a lot of data processing were more likely to be assigned to AI with a greater level of autonomy, due to its efficiency and accuracy. In addition, the **urgency** of a task was also taken into account. Using AI at a high level of autonomy means it is able to carry out the work around the clock, which is important as "*the decision-making time and decision-making paths are becoming ever shorter. Time is becoming a massive success factor [...]*" (ME9).
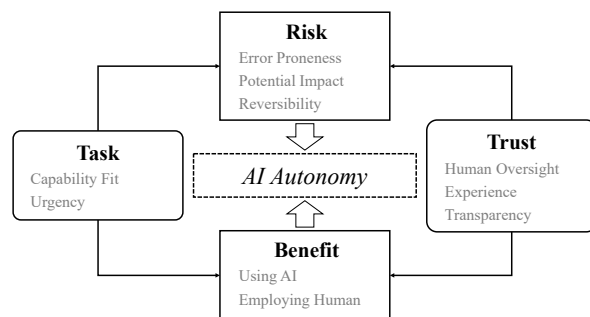
Figure 1: AI Autonomy Decision Framework

Trust: Experience and the importance of transparency. The experts' trust in AI played a major role in how autonomous they believed AI should be. We were able to identify three important aspects for trust in AI: human oversight, the experts' experience with AI, and AI transparency. A frequently articulated idea around **human oversight** was that "*in principle, I am fundamentally of the opinion that [AI] is a great support, but not with blind trust, never trust, and verify*" (ME8). This concept envisions an evolving process where humans closely monitor AI. However, as articulated by one expert, when "*you realize how good the output of the AI actually is*" (ME7), the necessity for human oversight diminishes. This shift is attributed to experts accumulating sufficient **experience** with AI, fostering a sense of trust. Additionally, comprehensive and sound regulation or trustworthy providers of AI models were mentioned as strengthening the trust in AI. Conversely, observing wrong AI predictions and experiencing misuse of AI for misinformation, propaganda, or social scoring, were detrimental to the expert's trust.

To better judge and incorporate AI results into their decisions, experts stressed the need for **transparent AI results** to understand the parameters leading to the AI's results, and to assess "*whether what the engine gives me is actually correct and has not been interpreted by the engine*" (ME12). Additionally, experts emphasized the importance of the **AI model's transparency** to understand how a model works, what data was used to train it, and what technical infrastructure it relies on. Uncertainties relating to the AI model's data processing and storage, or the potential accessibility of the data by malicious actors, were described as hindering the adoption by multiple experts.

Experts also stated a need for a deterministic tool that "*should always make the same decisions under the given parameters.*" (ME19), allowing them to rely on AI tools more. Overall, most experts expressed trusting AI, however many had conditions or requirements bound to their trust. Seven experts self-reported not to trust AI, as one put it "*I have relatively little trust in systems that act autonomously,[...] I really lack*

the basic trust that no unnecessary damage will be done."
(ME9).

Risks: Error-proneness, impact and reversibility. AI's
proneness to error, but also that of humans, played a central
role in the experts' assessment of risks related to a certain
level of autonomy. If AI were to be perfectly accurate in its
calculations, predictions and detection rates, then experts felt
it should be used for most cybersecurity tasks. Humans were
described to be fallible and generally not suitable for many
tasks in cybersecurity. Hence, AI without human influence
could strengthen cybersecurity, as AI was portrayed as a re-
liable tool that is not influenced by emotions nor biased like
humans might be.

Experts further determined the risk for a specific level of AI
autonomy also by the respective **impact**, i.e., the extent and
severity of the consequences, that an autonomous AI action
could cause. When assessing consequences, experts consid-
ered how far-reaching they were, such as whether they were
isolated to single workstations or could lead to company-wide
shutdowns. The more severe and serious these consequences
were perceived, the more conservatively experts were in grant-
ing higher levels of AI autonomy. Likewise, experts differed
notably in their judgement based on task domain. For instance,
for medical applications or critical infrastructure, experts were
reluctant to consider allowing high AI autonomy.

In addition, the **reversibility** of a particular actions played a
role. If the costs of a successful attack were high and the AI's
action reversible, experts described that they would rather let
AI react autonomous and too quickly and undo the action later
than hesitate too long. "*Whether the impact of the action can
be reversed and is fundamentally justifiable for the company*"
(ME8) generally plays an important role in the allocation of
autonomy. However, if the result is permanent, experts were
more reluctant to grant AI autonomy.

**Benefits: What can be achieved with AI, that can not be
done without.** When evaluating the benefits, experts com-
pared **the advantages** that are **gained when AI is used** to
advantages **gained when a human performs** a task. This
assessment is predominantly determined by the nature of the
task and the capability-task fit, as illustrated in Figure 1.

In the case that a task or an incident is time-sensitive, ex-
perts tended to lean towards giving AI more autonomy since
"*[t]ime is becoming a massive success factor in [...], in a
ransomware attack, the time between initial infection and
outbreak is sometimes less than an hour. And people have to
make sure they keep up.*" (ME9). They further believed "*that
the reliability of an automated solution is generally higher.*"
(ME5). In other words, this human approval can also be detri-
mental to the purpose.

Experts understood that "*if AI is perfectly trained to make
a quick decision in a critical case, it may be better than a
human being*" (CE24). Yet, experts also understood that the

outcome and consequences of AI acting autonomously needs
to be considered when deciding how much autonomy should
be given to AI.

Level of autonomy: in between risks and benefits. Based
on this framework of considerations (see Figure 1) some tasks
were frequently mentioned for specific levels of autonomy
and expert-AI task division. An additional tabular summary
can be found in Appendix F.

Level 1 - Decision support: Particularly in areas with a high
level of personal responsibility, or areas that require contex-
tual knowledge or creativity, the experts stated that they would
prefer to use AI only as a support for decision-making. This
enables them to offload routine tasks partially, while poten-
tially increasing their accuracy. In these cases, AI could sum-
marise information to support risk management, or generate
preliminary drafts for policies to support the experts in their
tasks and processes.

Level 2 - Human Approval: When tasks fit the AI's capabili-
ties, but are not time-critical but potentially have far-reaching
consequences, experts increasingly expressed the wish to have
AI propose a decision but still be able to review and ultimately
accept or reject the proposed decision. Exemplar tasks men-
tioned by the experts were the response actions to major
security incidents, including patching and the isolation of en-
tities.

Level 3 - Human Veto: Regarding most tasks with far-
reaching consequences or consequences that could not be
easily reversed, the experts expressed a desire to at least be
able to veto the actions of AI. While experts on the one hand
wanted the option to intervene, AI could still automatically
perform tasks without human interaction. However, experts
seemed to have difficulty understanding, or applying this level
to tasks that could be performed by AI.

Level 4 - Execute and Inform: In time-critical situations as
well as in situations in which the consequences of the AI's
actions could be mitigated or reversed, experts were willing to
let AI autonomously execute tasks but still expressed the wish
to be informed to maintain situational awareness. Exemplar
tasks were automated penetration testing or the automated
configuration of firewalls.

Level 5 - Fully automated: The experts mostly stated that
they were willing to delegate routine tasks or tasks with min-
imal impact to AI for fully autonomous processing. In such
cases, they would not want to be informed about each of the
AI's actions, as this could quickly lead to information over-
load, especially in the context of daily routine tasks. Exemplar
tasks were the distribution and verification of user privileges,
and continuous security monitoring.

4.3 Designing Security Expert-AI Interaction

This section outlines the experts' requirements and prefer-
ences for designing expert-AI interfaces. One way in which

experts described AI was a tool they could delegate time-consuming tasks to. A different, somewhat anthropomorphized view that the CSEs described was an AI assistant, where the AI's work is primarily supportive and at the request of the expert. Finally, several experts referred to AI as a co-pilot that actively thinks along and contributes to their shared work, as opposed to simply waiting to receive human input in order to perform tasks. Experts often considered themselves to be in a supervision role that evaluates, reviews and makes decisions with the support of AI. It was considered important for the AI tool to be available to the experts at all times and to be meaningfully integrated into their existing workflows. Experts emphasized the importance of seamlessly integrating AI into their existing processes, workflows, and routines so that day-to-day work is not complicated through additional steps. The communication between experts and AI needs to be appropriate to the situation. Many experts expressed the desire for communication through natural language. However, experts disagreed whether this should be through spoken word or via text. For important purposes, such as alerting, AI should address the expert proactively, whereas in other cases, experts wanted to be the ones initiating communication. The experts also urged that they must be able to integrate the context and strategy into the joint work when working on difficult cases. The experts therefore saw themselves in the role of the final decision-maker.

5   Discussion

In the following, we reflect on the limited deployment of AI tools in cybersecurity and factors that can enable successful integration. We discuss trust and its effects on experts' willingness to collaborate with AI and grant it specific levels of autonomy, the autonomy characteristics enabling expert-AI collaboration in cybersecurity. We conclude by highlighting the contributions of our AI autonomy framework for cybersecurity and provide design recommendations for expert-AI collaboration in cybersecurity.

Why is there little AI adoption?   CSEs expressed wanting and needing support in various areas of their jobs, including generating policies and awareness materials, facilitating low-stakes communication, supporting compliance with regulations, as well as monitoring, pattern detection, and even automated threat responses. Our results show that experts would appreciate support from AI and are willing to widely employ it in cybersecurity, but that the use of AI for a majority of daily tasks is still perceived as hypothetical. Even though AI tools already exist for most cybersecurity use cases, as elaborated in subsection 2.1, only one-third of organisations reported using, or planning the use of AI tools for protecting their digital assets [46, 92].

One reason for this misalignment could be a mismatch be-

tween the characteristics of the currently available solutions and the experts' considerations for deploying AI with a certain level of autonomy. For example, our findings indicated that aside from the type of task, trust in AI played an important role in deciding on AI autonomy and adoption. As usage experience was mentioned as a relevant factor for trust in AI tools, another reason could be a lack of experience with existing AI tools and a potential hesitance to be among the first to adopt these tools in a high-stakes environment. The following sections discuss potential reasons along with implications related to building trust in AI tools that may ultimately enhance adoption rates of existing solutions or inform the development of matching AI tools.

The importance of good experiences and usability.   The limited adoption of AI might also be related to negative experiences. Since experts stated to build trust primarily through experiences, encountering issues with AI systems can have a strong dissuading effect. This notion was reinforced by one expert, who reported a test with an AI tool in their organisation that turned out negatively, discouraging the expert from using AI in support of cybersecurity tasks. Other scientific studies confirm that if people can observe AI mistakes or malfunctions, the future outcomes of AI are viewed with more distrust [8], and its adoption becomes unlikely. Usability and reliability, at least in the early stages, impact the trust development with the introduced system [65], emphasizing the importance of these factors during system design. Also, in our interviews, experts placed a lot of importance on an initial familiarisation period. Conflicts during this initial exploratory phase could lead to an aversion to further, more sophisticated integration. AI tools, especially when used by domain experts, such as cybersecurity experts should be designed to be usable to that user group, and also prove to be reliable. Additionally, the quality of the AI tool should be carefully evaluated before deploying it, contributing to the secure integration of AI tools into an organisation, and at the same time also the trust building processes between experts and AI.

Building trust with the AI "employee".   While experts almost unanimously described AI as a tool for their work, their descriptions of trust-building were closely related and sometimes even narrated using the example of a subordinate employee. In line with that, previous research has hypothesized that building trust between humans and AI seems to mirror processes of human-human relationships [34, 53]. This unintentional anthropomorphisation of the AI tools could indicate the experts' desire to relate with the AI tool and impact the experts' intention to use AI agents, but not directly induce trust [17, 98]. Experts repeatedly described that they would build trust in an AI tool by first giving it less important tasks, at a low level of autonomy, and closely monitoring its performance. Consistent with previous findings [1], experts appeared willing to give the system more freedom once the AI

tool has proven to produce reliable results. Then, they might deploy a higher level of autonomy and use the AI for more significant tasks. Observing the trustee's performance, and their reliability were found to be correlated with the trustworthiness of a subordinate, and prior research already suggested that this might also apply to human-machine relations [34].

I should "never trust, and verify", or should I?   Regardless of trust-building processes, for many experts trust in AI was still based on the possibility to oversee its work (see Figure 1). As experts have a high level of self-confidence and expertise in the domain where AI would be deployed, they could rely on their own capabilities to verify and correct AI results. This "never trust, and verify" principle, as one expert called it, strongly relies on the human as a final decision power to validate, verify and modify the AI outcome and to prevent damage or negative consequences. Especially, with the high availability of LLMs, the potential of AI to err has become even more visible to the user through AI hallucinations [6, 78], further increasing their desire to oversee the outcomes of AI models and validate their correctness. While manual verification is a good approach to mitigate negative consequences originating from the use of AI, it can also hinder the advantages gained through the collaboration of experts and AI. If experts always had the final say, AI systems would not only be inhibited in relieving experts of repetitive tasks due to requiring manual approval for actions but could even end up unable to react in real-time in critical situations. Lee and See [53] already pointed out that people with high self-confidence and low trust in automation tend to fall back to manual control more quickly, diminishing the benefits that AI could provide. To be able to leverage the capabilities of automation and intelligent systems in cybersecurity, it is therefore important that experts trust AI enough to the point where they do not feel a need to manually validate and verify all AI actions. Thus, the interaction of autonomy levels and the factors influencing trust need to be researched further to be able to design trustworthy and therefore effective AI systems for cybersecurity.

Influence of AI autonomy: considerations on AI adoption.   At this point, we reflect on the experts' requirements for cybersecurity and how these relate to the four autonomy characteristics - adaptability, transparency, determinism, and openness - introduced by Simmler and Frischknecht [80] (see Table 2). First, the **adaptability** of the system was not so much a requirement for AI in cybersecurity expressed by the experts, but more so assumed by them to be one of the unique capabilities of AI as a technology. **Transparency** proved to be a relevant aspect influencing trust in AI, as shown in the AI autonomy decision framework in Figure 1. With the term transparency, experts mostly described the need for transparency and understandability of AI outputs, which is crucial for making quick yet well-founded decisions. The need for AI tools

to be transparent arises from the complex and high-stakes nature of the cybersecurity environment. If discrepancies in judgment between the experts and AI arise, experts must be able to understand the factors that led to this discrepancy, allowing them to assess and correct the collaborative output based on their expertise. Similarly, Vössing et al. [97] argued that the ability to correct the output of AI, and exercise control can build trust. They found that providing explanations on the AI models' reasoning for collaborative task solving strengthened the cognition-based trust and reduced the discrepancies between the human mental model and the AI's embedded decision model, contributing to a successful collaboration [97]. Especially in the cybersecurity context, experts need to additionally understand the AI's reasoning to assess AI tools for potential tampering by malicious actors. Leveraging methods of explainable AI to provide transparency and information on the AI's reasoning therefore is a promising approach to improve collaboration of experts and AI in cybersecurity.

The need for transparency and sound reasoning in high-stakes decisions is not only preferred by experts but also a requirement by cybersecurity regulations [96], and an important factor in the EU AI Act for AI tools that humans interact with [27]. While the primary purpose of AI transparency was to gain trust, experts also described their need for results they can justify and comprehend in their responsibility towards management and legal authorities.

Further, experts need to be able to rely on AI tools that always react the same way given the same input for some critical cases, requiring a **deterministic** system. Such determinism can additionally increase the experts' perception of the AI's reliability, as it ensures predictability. Reliability additionally is an important aspect of trust [34], indicating that a semi-deterministic system could foster trust between experts and AI.

Many experts expressed worries about the potential to extract sensitive data if the AI tool was connected to additional sources or the internet, leading to the desire for transparency of AI models where they could at least partially quantify this risk and observe how the model handles, stores and distributes the data that users put in. An AI tool that could be deployed in cybersecurity needs to be **closed** [80] as the experts did not see the possibility of mitigating such risks otherwise. Crucially, this is important for even minor tasks like communication or summarising documents, as regulations and policies prohibit information from being shared. Therefore, any open system is not a viable option to enhance an organisation's cybersecurity, as the possibilities to tamper with AI models are still not yet fully understood nor mitigable.

In addition to the 'closed-loop' factor, it is also important to consider that feasibility also plays a role in the practical application of AI. Despite the digital nature of cybersecurity, most data still needs to be processed or even digitised in order to be useful for AI, which can further limit the practical feasibility for individual use cases. To wrap up our discussion and as a

final conclusion, we compare our autonomy framework with existing frameworks in the literature. This should highlight differences, but also similarities, to show consistency with other literature on the one hand, but also uniqueness for the collaboration of cybersecurity experts with AI.

Comparison of autonomy-focused human-AI collaboration frameworks. While Salikutluk et al.'s [75] framework includes self-confidence as a fundamental factor, this factor is less relevant to our target population of CSEs, as it will always be high, and is therefore not represented in our framework. The effects of task failure and competence comparison are similar to our framework's risk and capability fit, respectively. As Salikutluk et al.'s framework draws from psychology, it includes theory of mind as an important factor that describes the understanding and awareness of the capabilities of different actors. This factor is included in our model through the notion of transparency, the CSEs' desire for transparent AI outcomes and models aligns with the need to understand and be aware of the other actor's capabilities [75]. While their model follows a flat structure, where all factors directly influence autonomy, our decision framework has two tiers, where the underlying task and trust factors then influence a risk and benefit trade-off. This difference might stem from CSEs having a more transactional view guided by higher-level cost versus reward considerations. CSEs are frequently exposed to risk-benefit analysis, and might naturally fall back to similar mechanisms for assessments of AI tools' autonomy. Our model is also not primarily developed under the assumption of a shared physical workspace setting; therefore, less emphasis is put on whether a change of human action is required. CSEs did emphasize the need for suitable integration of AI agents into their already existing workflows, making the change of human action undesirable. While both frameworks are established in different ways, ours through expert interviews imagining theoretical applications, and theirs with an emulated shared workspace with a physical AI agent, the similarities highlight the importance of some factors, shared between modalities. In particular, the capability fit for a task and the risk posed by failure seems to be perceived as important by humans when interacting with an AI system in either scenario. Future research should deepen the understanding of how the interaction medium and human expertise, and thereby self-confidence, affect the degree of autonomy afforded to AI systems.

5.1 Limitations & Future Work

Like all research, this study is subject to several limitations. *First,* our qualitative approach does not allow for quantification of the findings and can thus be viewed as an initial step towards informing future (quantitative) research on expert-AI collaboration in cybersecurity. *Second,* the sample was mostly male security professionals. This gender imbalance is not desired, but representative of the target group, as women are

underrepresented in cybersecurity [42]. The sample included mainly experts with strategic and managing tasks. Future work could extend our high-level insights into cybersecurity professionals on operational levels. *Third,* we employed AI-adapted versions of the GAToRS [51] and the Human-Computer Trust scale [54] to better understand the participants' attitudes towards AI. However, it was challenging to identify meaningful quantitative patterns related to the multi-faceted qualitative data. Future work could evaluate these scales for AI-related research in quantitative settings with larger samples.

5.2 Summary: Recommendations for Expert-AI Collaboration in Cybersecurity

In sum, AI tools designed for effective collaboration with CSEs must address the demands of complexity, uncertainty, and high stakes in the field of cybersecurity. AI tools need to fulfil the high requirements related to data security and transparency and enable CSEs to make meaningful disclosures to management and legal authorities. At the same time, to leverage the complementary capabilities of experts and AI in cybersecurity, further understanding the effects of varying degrees of AI autonomy on expert trust is crucial for long-term adoption and successful collaboration. Therefore, the tool needs to have the following characteristics:

- the AI output needs to be transparent and understandable for CSEs, as well as the AI model and its respective infrastructure,

- the AI tool needs to be designed to accommodate the process of building trust allowing for low to high autonomy levels,

- the AI model needs to be closed to protect the data it is processing, semi-deterministic to accommodate for known best practices, but also adaptable to new threats to accommodate for the quickly evolving threat landscape of cybersecurity.

Data Availability Statement

Due to the high sensitivity of interviews with regard to the potential identification of participants through AI tools, the interview data is not openly available. Detailed sample information, the interview guide, codebook, and exemplary quotes are provided in the article and Appendix. For further information or access to the original interview transcripts, please contact the authors.

References

[1] Barbara D Adams, Lora E Bruyn, Sébastien Houde, Paul Angelopoulos, Kim Iwasa-Madge, and Carol Mc-

Cann. Trust in automated systems. Technical report, Ministry of National Defence, 2003.

[2] M.E. Ahsen, M.U.S. Ayvaci, and R. Mookerjee. When Machines Will Take Over? Algorithms for Human-Machine Collaborative Decision Making in Healthcare. In *Proceedings of the 56th Hawaii International Conference on System Sciences*, volume 202, pages 5733–5740, 2023.

[3] U. Aickelin, M. Maadi, and H.A. Khorshidi. Expert-Machine Collaborative Decision Making: We Need Healthy Competition. *IEEE Intelligent Systems*, 37(5):28–31, 2022.

[4] Matt Aiello, Scott Thompson, Max Randria, Camilla Reventlow, Guy Shaul, and Adam Vaughan. 2022 Global Chief Information Security Officer (CISO) Survey - Insights - Heidrick & Struggles. Technical report, Heidrick & Struggles, 2022.

[5] Khalid Al-Rowaily, Muhammad Abulaish, Nur Al-Hasan Haldar, and Majed Al-Rubaian. BiSAL – A bilingual sentiment analysis lexicon to analyze Dark Web forums for cyber security. *Digital Investigation*, 14:53–62, September 2015.

[6] Athaluri Sai Anirudh, Manthena Sandeep Varma, Kesapragada V. S. R. Krishna Manoj, Yarlagadda Vineel, Dave Tirth, and Duddumpudi Rama Tulasi Siri. Exploring the Boundaries of Reality: Investigating the Phenomenon of Artificial Intelligence Hallucination in Scientific Writing Through ChatGPT References. *Cureus*, 15(4), 2023.

[7] American Psychological Association et al. Ethical principles of psychologists and code of conduct. Technical report, American Psychological Association, 2016. Retrieved 14th February 2024 from https://www.apa.org/ethics/code.

[8] Elizabeth Bondi, Raphael Koster, Hannah Sheahan, Martin Chadwick, Yoram Bachrach, Taylan Cemgil, Ulrich Paquet, and Krishnamurthy Dvijotham. Role of Human-AI Interaction in Selective Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(5):5286–5294, June 2022.

[9] Adriana Braga and Robert K. Logan. The Emperor of Strong AI Has No Clothes: Limits to Artificial Intelligence. *Information*, 8(4):156, December 2017.

[10] Chuck Brooks. Cybersecurity Trends & Statistics For 2023; What You Need To Know, May 2023. Forbes.

[11] Chuck Brooks and Frederic Lemieux. Three Key Artificial Intelligence Applications For Cybersecurity by Chuck Brooks and Dr. Frederic Lemieux, September 2021. Forbes.

[12] Jason W. Burton, Mari-Klara Stein, and Tina Blegind Jensen. A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, 33(2):220–239, 2020.

[13] Federico Cabitza, Andrea Campagner, Luca Ronzio, Matteo Cameli, Giulia Elena Mandoli, Maria Concetta Pastore, Luca Maria Sconfienza, Duarte Folgado, Marília Barandas, and Hugo Gamboa. Rams, hounds and white boxes: Investigating human-AI collaboration protocols in medical diagnosis. *Artificial Intelligence In Medicine*, 138, April 2023.

[14] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019.

[15] Capgemini Research Institute. AI in Cybersecurity. https://www.capgemini.com/news/press-releases/ai-in-cybersecurity/, July 2019.

[16] Miguel V. Carriegos, Ángel L. Muñoz Castañeda, M. T. Trobajo, and Diego Asterio De Zaballa. On Aggregation and Prediction of Cybersecurity Incident Reports. *IEEE Access*, 9:102636–102648, 2021.

[17] Qian Qian Chen and Hyun Jung Park. How anthropomorphism affects trust in intelligent personal assistants. *Industrial Management & Data Systems*, 121(12):2722–2737, January 2021.

[18] Hyesun Choung, Prabu David, and Arun Ross. Trust in AI and Its Role in the Acceptance of AI Technologies. *International Journal of Human–Computer Interaction*, 39(9):1727–1739, May 2023. arXiv:2203.12687 [cs].

[19] Ylona Chun Tie, Melanie Birks, and Karen Francis. Grounded theory research: A design framework for novice researchers. *SAGE Open Medicine*, 7, January 2019.

[20] Adele E Clarke. Grounded theory: Critiques, debates, and situational analysis. *The SAGE Handbook of Social Science Methodology*, pages 423–442, 2007.

[21] Juliet M. Corbin and Anselm Strauss. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology*, 13(1):3–21, March 1990.

[22] Darktrace. AI Cyber Security Solutions. https://www.darktrace.com, 2024. Accessed: 2024-06-06.

[23] Jack H Davenport. Collaborative human-machine analysis to disambiguate entities in unstructured text and structured datasets. In *Next-Generation Analyst IV*, volume 9851, pages 16–22. SPIE, 2016.

[24] Thomas H Davenport and Rajeev Ronanki. Artificial intelligence for the real world: Don't start with moon shots. *Harvard business review*, 96(1):108–116, 2018.

[25] Dominik Dellermann, Philipp Ebel, Matthias Söllner, and Jan Marco Leimeister. Hybrid Intelligence. *Business & Information Systems Engineering*, 61(5):637–643, October 2019.

[26] Mica R Endsley. The application of human factors to the development of expert systems for advanced cockpits. In *Proceedings of the Human Factors Society Annual Meeting*, volume 31, pages 1388–1392. SAGE Publications, Los Angeles, CA, 1987.

[27] European Union. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206, 2021. COM/2021/206 final.

[28] Zhen Fang, Xinyi Zhao, Qiang Wei, Guoqing Chen, Yong Zhang, Chunxiao Xing, Weifeng Li, and Hsinchun Chen. Exploring key hackers and cybersecurity threats in Chinese hacker communities. In *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, pages 13–18, September 2016.

[29] Alessandro Fausto, Giovanni Battista Gaggero, Fabio Patrone, Paola Girdinio, and Mario Marchese. Toward the Integration of Cyber and Physical Security Monitoring Systems for Critical Infrastructures. *Sensors*, 21(21):6970, January 2021.

[30] Lorenzo Fernández Maimó, Alberto Huertas Celdrán, Ángel L. Perales Gómez, Félix J. García Clemente, James Weimer, and Insup Lee. Intelligent and Dynamic Ransomware Spread Detection and Mitigation in Integrated Clinical Environments. *Sensors*, 19(5):1114, January 2019.

[31] David Gefen, Izak Benbasat, and Paula Pavlou. A Research Agenda for Trust in Online Environments. *Journal of Management Information Systems*, 24(4):275–286, April 2008.

[32] Shirley Gregor and Izak Benbasat. Explanations from Intelligent Systems: Theoretical Foundations and Implications for Practice. *MIS Quarterly*, 23(4):497–530, 1999. Publisher: Management Information Systems Research Center, University of Minnesota.

[33] P. A. Hancock. Politechnology: Manners Maketh Machine. *Human-Computer Etiquette: Cultural Expectations and the Design Implications They Place on Computers and Technology*, January 2010.

[34] P. A. Hancock, Theresa T. Kessler, Alexandra D. Kaplan, Kimberly Stowers, J. Christopher Brill, Deborah R. Billings, Kristin E. Schaefer, and James L. Szalma. How and why humans trust: A meta-analysis and elaborated model. *Frontiers in Psychology*, 14:1081086, March 2023.

[35] Patrick Hemmer, Max Schemmer, Michael Vössing, and Niklas Kühl. Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review. In *PACIS 2021 Proceedings*, 2021.

[36] Patrick Hemmer, Monika Westphal, Max Schemmer, Sebastian Vetter, Michael Vössing, and Gerhard Satzger. Human-AI Collaboration: The Effect of AI Delegation on Human Task Performance and Task Satisfaction. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, pages 453–463, New York, NY, USA, March 2023. Association for Computing Machinery.

[37] Silvana Hinsen, Peter Hofmann, Jan Jöhnk, and Nils Urbach. How can organizations design purposeful human-AI interactions: A practical perspective from existing use cases and interviews. In *Hawaii International Conference on System Sciences*, 2022.

[38] Martin Husák. Towards a Data-Driven Recommender System for Handling Ransomware and Similar Incidents. In *2021 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 1–6, November 2021.

[39] IBM. IBM Security Guardium. https://www.ibm.com/security/data-security/guardium, Dec 2018. Accessed: 2024-06-06.

[40] IBM. IBM Security Verify. https://www.ibm.com/products/verify-saas, 2024. Accessed: 2024-06-06.

[41] ISACA. State of cybersecurity 2023 report, 2023. Retrieved 9th June 2024 from: https://www.isaca.org/resources/reports/state-of-cybersecurity-2023.

[42] (ISC)2. Cybersecurity Workforce Study, 2023. Retrieved 14th February 2024 from: https://www.isc2.org/research.

[43] Shintaro Ishikawa, Seiichi Ozawa, and Tao Ban. Port-Piece Embedding for Darknet Traffic Features and Clustering of Scan Attacks. In Haiqin Yang, Kitsuchart Pasupa, Andrew Chi-Sing Leung, James T. Kwok, Jonathan H. Chan, and Irwin King, editors, *Neural Information Processing*, Lecture Notes in Computer Science, pages 593–603, Cham, Switzerland, 2020. Springer International Publishing.

[44] Mohammad Hossein Jarrahi. Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons*, 61(4):577–586, July 2018.

[45] Kaspersky. Kaspersky Endpoint Security for Business. https://www.kaspersky.com/enterprise-security/endpoint, 2024. Accessed: 2024-06-06.

[46] Ramanpreet Kaur, Dušan Gabrijelčič, and Tomaž Klobučar. Artificial intelligence for cybersecurity: Literature review and future research directions. *Information Fusion*, 97:101804, September 2023.

[47] Ujwal Kayande, Arnaud De Bruyn, Gary L. Lilien, Arvind Rangaswamy, and Gerrit H. van Bruggen. How Incorporating Feedback Mechanisms in a DSS Affects DSS Evaluations. *Information Systems Research*, 20(4):527–546, December 2009.

[48] Gary A. Klein. *Sources of Power, 20th Anniversary Edition: How People Make Decisions*. MIT Press, September 2017.

[49] Hansaka Angel Dias Edirisinghe Kodituwakku, Alex Keller, and Jens Gregor. InSight2: A Modular Visual Analysis Platform for Network Situational Awareness in Large-Scale Networks. *Electronics*, 9(10):1747, October 2020. Number: 10 Publisher: Multidisciplinary Digital Publishing Institute.

[50] J. E. Korteling, G. C. van de Boer-Visschedijk, R. A. M. Blankendaal, R. C. Boonekamp, and A. R. Eikelboom. Human- versus Artificial Intelligence. *Frontiers in Artificial Intelligence*, 4, 2021.

[51] Mika Koverola, Anton Kunnari, Jukka Sundvall, and Michael Laakasuo. General Attitudes Towards Robots Scale (GAToRS): A New Instrument for Social Surveys. *International Journal of Social Robotics*, 14(7):1559–1581, September 2022.

[52] Chang-Eun Lee, Jaeuk Baek, Jeany Son, and Young-Guk Ha. Deep AI military staff: cooperative battlefield situation awareness for commander's decision making. *Journal Of Supercomputing*, 79(6):6040–6069, April 2023.

[53] John D. Lee and Katrina A. See. Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, 46(1):50–80, 2004.

[54] Maria Madsen and Shirley Gregor. Measuring human-computer trust. In *11th australasian conference on information systems*, volume 53, pages 6–8. Citeseer, 2000.

[55] Malwarebytes. Malwarebytes. https://www.malwarebytes.com, 2024. Accessed: 2024-06-06.

[56] Daniel L. Marino, Chathurika S. Wickramasinghe, Billy Tsouvalas, Craig Rieger, and Milos Manic. Data-Driven Correlation of Cyber and Physical Anomalies for Holistic System Health Monitoring. *IEEE Access*, 9:163138–163150, 2021.

[57] D. Harrison Mcknight, Michelle Carter, Jason Bennett Thatcher, and Paul F. Clay. Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on Management Information Systems*, 2(2):12:1–12:25, July 2011.

[58] Merriam-Webster. autonomy. Retrieved 16th February 2024 from https://www.merriam-webster.com/dictionary/autonomy.

[59] Christian Meske and Enrico Bunde. Transparency and Trust in Human-AI-Interaction: The Role of Model-Agnostic Explanations in Computer Vision-Based Decision Support. In Helmut Degen and Lauren Reinerman-Jones, editors, *Artificial Intelligence in HCI*, volume 12217, pages 54–69. Springer International Publishing, Cham, 2020.

[60] Benjamin S. Meyers and Andrew Meneely. An Automated Post-Mortem Analysis of Vulnerability Relationships using Natural Language Word Embeddings. *Procedia Computer Science*, 184:953–958, January 2021.

[61] Microsoft. Microsoft erklärt: Was ist künstliche Intelligenz? Definition & Funktionen von AI, March 2020.

[62] Microsoft. Microsoft Security Copilot. https://securitycopilot.microsoft.com/, 2024. Accessed: 2024-06-06.

[63] Ali R. Montazemi. The impact of experience on the design of user interface. *International Journal of Man-Machine Studies*, 34(5):731–749, May 1991.

[64] Sridhar Muppidi, Lisa Fisher, and Gerald Parham. AI and automation for cybersecurity. Benchmark, IBM Institute for Business Value, United States of America, June 2022.

[65] Lea S. Müller, Christoph Nohe, Sebastian Reiners, Jörg Becker, and Guido Hertel. Adopting information systems at work: a longitudinal examination of trust dynamics, antecedents, and outcomes. *Behaviour & Information Technology*, 43(6):1096–1128, April 2024.

[66] Scott Brave Nass, Cliff. Emotion In Human-Computer Interaction. In *The Human-Computer Interaction Handbook*. CRC Press, 2 edition, 2007.

[67] Pantaleone Nespoli, Félix Gómez Mármol, and Jorge Maestre Vidal. A Bio-Inspired Reaction Against Cyberattacks: AIS-Powered Optimal Countermeasures Selection. *IEEE Access*, 9:60971–60996, 2021.

[68] Yannis Nikoloudakis, Ioannis Kefaloukos, Stylianos Klados, Spyros Panagiotakis, Evangelos Pallis, Charalabos Skianis, and Evangelos K. Markakis. Towards a Machine Learning Based Situational Awareness Framework for Cybersecurity: An SDN Implementation. *Sensors*, 21(14):4939, January 2021.

[69] Ingrid Nunes and Dietmar Jannach. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction*, 27(3):393–444, December 2017.

[70] Marc Pinski, Martin Adam, and Alexander Benlian. AI Knowledge: Improving AI Delegation through Human Enablement. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pages 1–17, New York, NY, USA, April 2023. Association for Computing Machinery.

[71] Yuanqing Qin, Yuan Peng, Kaixing Huang, Chunjie Zhou, and Yu-Chu Tian. Association Analysis-Based Cybersecurity Risk Assessment for Industrial Control Systems. *IEEE Systems Journal*, 15(1):1423–1432, March 2021.

[72] Panagiotis Radoglou-Grammatikis, Panagiotis Sarigiannidis, Eider Iturbe, Erkuden Rios, Saturnino Martinez, Antonios Sarigiannidis, Georgios Eftathopoulos, Yannis Spyridis, Achilleas Sesis, Nikolaos Vakakis, Dimitrios Tzovaras, Emmanouil Kafetzakis, Ioannis Giannoulakis, Michalis Tzifas, Alkiviadis Giannakoulias, Michail Angelopoulos, and Francisco Ramos. SPEAR SIEM: A Security Information and Event Management system for the Smart Grid. *Computer Networks*, 193:108008, July 2021.

[73] Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R. Varshney, Amit Dhurandhar, and Richard Tomsett. Deciding Fast and Slow: The Role of Cognitive Biases in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):83:1–83:22, April 2022.

[74] Jacob Sakhnini, Hadis Karimipour, Ali Dehghantanha, and Reza M. Parizi. Physical layer attack identification and localization in cyber–physical grid: An ensemble deep learning based approach. *Physical Communication*, 47:101394, August 2021.

[75] Vildan Salikutluk, Janik Schöpper, Franziska Herbert, Katrin Scheuermann, Eric Frodl, Dirk Balfanz, Frank Jäkel, and Dorothea Koert. An evaluation of situational autonomy for human-AI collaboration in a shared workspace setting. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, pages 1–17, New York, NY, USA, May 2024. Association for Computing Machinery.

[76] José Carlos Sancho, Andrés Caro, Mar Ávila, and Alberto Bravo. New approach for threat classification and security risk estimations based on security event management. *Future Generation Computer Systems*, 113:488–505, December 2020.

[77] Iqbal H. Sarker, A. S. M. Kayes, Shahriar Badsha, Hamed Alqahtani, Paul Watters, and Alex Ng. Cybersecurity data science: an overview from machine learning perspective. *Journal of Big Data*, 7(1):41, July 2020.

[78] Yiqiu Shen, Laura Heacock, Jonathan Elias, Keith Hentel, Beatriu Reig, George Shih, and Linda Moy. Chat-GPT and Other Large Language Models Are Double-edged Swords. *Radiology*, 307(2):e230163, April 2023.

[79] Dominik Siemon. Elaborating team roles for artificial intelligence-based teammates in human-ai collaboration. *Group Decision and Negotiation*, 31(5):871–912, 2022.

[80] Monika Simmler and Ruth Frischknecht. A taxonomy of human–machine collaboration: capturing automation and technical autonomy. *AI & SOCIETY*, 36(1):239–250, March 2021.

[81] Matthias Söllner, Axel Hoffmann, Holger Hoffmann, Arno Wacker, and Jan Marco Leimeister. Understanding the Formation of Trust in IT Artifacts. In *Proceedings of the International Conference on Information Systems, ICIS 2012*, volume 11, Orlando, December 2012. Association for Information Systems.

[82] Matthias Söllner, Axel Hoffmann, and Jan Marco Leimeister. Why different trust relationships matter for information systems users. *European Journal of Information Systems*, 25(3):274–287, May 2016.

[83] Richard Starnes, Sumit Cherian, and Luis Delabarre. Reinventing cybersecurity with artificial intelligence: The new frontier in digital security, July 2019. Retrieved 16th February 2024 from https://www.capgemini.com/insights/research-library/.

[84] Anselm L. Strauss and Juliet M. Corbin. *Basics of qualitative research: techniques and procedures for developing grounded theory*. Sage Publications, Thousand Oaks, 2. edition, 2003.

[85] S Shyam Sundar. Rise of machine agency: A framework for studying the psychology of human–AI interaction (HAII). *Journal of Computer-Mediated Communication*, 25(1):74–88, March 2020.

[86] Matthias Söllner and Paul Pavlou. A longitudinal perspective on trust in it artefacts. *Research Papers*, June 2016.

[87] Tenable, Inc. Tenable One Exposure Management Platform. https://www.tenable.com/products/tenable-one, October 2022. Accessed: 2024-06-06.

[88] Tessian. Complete Cloud Security Email Platform. https://www.tessian.com, 2024. Accessed: 2024-06-06.

[89] Zhiyi Tian, Lei Cui, Jie Liang, and Shui Yu. A Comprehensive Survey on Poisoning Attacks and Countermeasures in Machine Learning. *ACM Computing Surveys*, 55(8):166:1–166:35, December 2022.

[90] Trint Limited. Trint. https://trint.com/, 2022.

[91] Agnieszka A. Tubis, Sylwia Werbińska-Wojciechowska, Mateusz Góralczyk, Adam Wróblewski, and Bartłomiej Ziętek. Cyber-Attacks Risk Analysis Method for Different Levels of Automation of Mining Processes in Mines Based on Fuzzy Theory Use. *Sensors*, 20(24):7210, January 2020.

[92] TÜV-Verband. Einsatz künstlicher Intelligenz in der IT-Sicherheit in deutschen Unternehmen 2020. https://de.statista.com/statistik/daten/studie/1251115/umfrage/ki-in-der-it-sicherheit-in-unternehmen/, 2020.

[93] Takane Ueno, Yuto Sawa, Yeongdae Kim, Jacqueline Urakami, Hiroki Oura, and Katie Seaborn. Trust in human-AI interaction: Scoping out models, measures, and methods. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–7. ACM, April 2022.

[94] Erik Veitch and Ole Andreas Alsos. A systematic review of human-AI interaction in autonomous ship systems. *Safety Science*, 152:105778, August 2022.

[95] VERBI Software. MAXQDA, 2024. https://www.maxqda.com/.

[96] Javier Verdugo and Moisés Rodríguez. Assessing data cybersecurity using ISO/IEC 25012. *Software Quality Journal*, 28(3):965–985, September 2020.

[97] Michael Vössing, Niklas Kühl, Matteo Lind, and Gerhard Satzger. Designing Transparency for Effective Human-AI Collaboration. *Information Systems Frontiers*, 24(3):877–895, June 2022.

[98] Katja Wagner, Frederic Nimmermann, and Hanna Schramm-Klein. Is It Human? The Role of Anthropomorphism as a Driver for the Successful Acceptance of Digital Voice Assistants. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

[99] Johannes Weyer. Die Kooperation menschlicher Akteure und nicht-menschlicher Agenten: Ansatzpunkte einer Soziologie hybrider Systeme. Technical report, Wirtschafts- und Sozialwissenschaftliche Fakultät Universität Dortmund, 2006. Working Paper.

[100] Stacey M Whitecotton. The effects of experience and confidence on decision aid reliance: A causal model. *Behavioral Research in Accounting*, 8:194–216, 1996.

[101] Fan Zhang, Hansaka Angel Dias Edirisinghe Kodituwakku, J. Wesley Hines, and Jamie Coble. Multi-layer Data-Driven Cyber-Attack Detection System for Industrial Control Systems Based on Network, System, and Process Data. *IEEE Transactions on Industrial Informatics*, 15(7):4362–4369, July 2019.

[102] Zscaler, Inc. Zscaler Data Protection. https://www.zscaler.com/solutions/security-transformation/data-protection, 2024. Accessed: 2024-06-06.

Appendix

Appendix A: Automation/Autonomy Taxonomy

The following tables 3 and 2 detail the levels of automation and the dimensions of autonomy as described by [80].

Dimension	Description
Transparency	Degree to which all execution steps between an input A and an output Be are specified and transparent
Determinism	Degree to which an A always equally leads to an output B
Adaptability	Degree to which a system can learn and adapt behaviour to changing environments
Openness	Degree to which the system can expand its original input for collaboration and interaction

Table 2: Dimensions of autonomy based on Simmler & Frischknecht [80].

Level	Description	Explanation
1	Offers Decision	System makes recommendations, operator selects and decides
2	Executes with human approval	System makes recommendations and selects "best" option, operator (dis-)approves
3	Executes if no human vetoes	System makes recommendation, selects "best" option and executes, operator can correct and veto
4	Executes and then informs	System makes recommendation, selects "best" option, executes, and informs operator (passive operator role)
5	Executes fully automated	System makes recommendation, selects "best" option and executes without informing (operator not part of process)

Table 3: Levels of automation as described by Simmler & Frischknecht [80] and based on Endsley [26] and Weyer [99].

Appendix B: Complete Codebook

The complete codebook, including descriptions and examples for each code, have been published on the ETH Research Collection and are accessible via the following DOI: https://doi.org/10.3929/ethz-b-000674517.

Appendix C: AI-adapted General Attitudes Towards Robots Scale (GAToRS) [51]

No.	AI-adapted Item description
Personal Level Positive Attitude (P+)	
RA1	I can trust persons and organizations related to development of AI
RA2	Persons and organizations related to development of AI will consider the needs, thoughts and feelings of their users
RA3	I can trust in AI
RA4	I would feel relaxed interacting with an AI
RA5	If AI had emotions, I would be able to befriend them
Personal Level Negative Attitude (P-)	
RA6	I would feel uneasy if I was given a job where I had to use AI
RA7	I fear that an AI would not understand my commands
RA8	AI scares me
RA9	I would feel very nervous just being around an AI
RA10	I don't want an AI to talk to me
Societal Level Positive Attitude (S+)	
RA11	AI is necessary because it can do jobs that are too hard or too dangerous for people
RA12	AIs can make life easier
RA13	Assigning routine tasks to AIs lets people do more meaningful tasks
RA14	Dangerous tasks should primarily be given to AI
RA15	AI is a good thing for society because it helps people
Societal Level Negative Attitude (S-)	
RA16	AI may make us even lazier
RA17	Widespread use of AI is going to take away jobs from people
RA18	I am afraid that AI will encourage less interaction between humans
RA19	AI is one of the areas of technology that needs to be closely monitored
RA20	Unregulated use of AI can lead to societal upheavals
Criterion Items	
C1	Generally speaking, I have a positive view of AI
C2	I have personal experience of using AI
C3	I am interested in scientific discoveries and technological developments
C4	AI is a familiar topic to me

Table 4: GAToRS by [51] and adapted to the AI use case.

Sub Scale	Min	Mean	Max	SD
Personal+	13.00	19.81	24.00	3.12
Personal-	7.00	12.19	19.00	3.01
Societal+	18.00	25.15	31.00	3.06
Societal-	11.00	23.19	28.00	4.11

Table 5: Sample scores of the AI-adapted GAToR Scale [51], $n = 27$

Appendix D: Interview guideline

The following section describes the interview guideline consisting of three focus areas: 1) Understanding the job, 2) Understanding the type or level of human-AI collaboration, and 3) Understanding the hopes, fears, and emotions.

Focus 1: Understanding the job.

RQ: What tasks do cybersecurity experts need to complete, and which could hypothetically be automated or complemented by AI?

Preface: Talk about role in general, not considering AI at this point.

Role description

What is your current role in your organization?

What are your key responsibilities in your role?

Ask if not already mentioned: Which tasks do you need to do in order to achieve your goal(s) and fulfil your key responsibilities?

Need for support

For which of the tasks are you lacking resources, e.g., time or skills?

In case not answered: Which tasks would you like support from an automation or AI solution for?

Focus 2: Understanding the type or level of human-AI collaboration.

RQ: What is the expert's view on the feasibility of integrating AI into their workflow to collaborate?

Reflecting on Human Capabilities

What do you imagine AI is generally good at doing?

Which of the tasks are well suited for AI in the domain of cybersecurity?

Reflecting on AI Capabilities

What do you imagine AI is generally good at doing?

Which of the tasks are well suited for AI in the domain of cybersecurity?

Collaboration of Humans and AI

What would the interaction between you as an expert and the AI be?

Feasibility

We have talked about which tasks are well suited, but for which would you like to integrate AI and why?

Where would you hesitate to use AI?

What kind of support do you require? E.g., decision-aid (i.e., information gathering) or automated responses (i.e., AI can react to detected threats itself) etc. Why?

Feasibility *(continued)*

What limitations might arise from AI as a technology?

In practice, what might be other factors limiting or impacting the feasibility of the human-AI collaboration?

Having discussed the tasks and aspects of their feasibility now, I would like to ask you to sort the tasks into this how-now-wow matrix (visualized in Figure 2). The tasks and collaboration ideas you think feasible and find somewhat desirable should go into now. The tasks and collaboration ideas you find desirable but are unsure of the feasibility should go into the how, and ideas that are highly desirable and at the same time highly feasible, please place into the wow matrix.

Focus 3: Understanding the hopes, fears, and emotions.

(Former) RQ: Which emotions, hopes, and fears do expert-AI collaboration trigger in experts, and for which reasons?

Trust

Could you, as one half of this collaboration, trust the AI?

What do you need to establish and maintain trust in this technology and the collaboration?

What could cause you to lose trust in AI?

Hopes

What are your hopes considering expert-AI collaboration?

Worries

What are your worries about the integration of AI into your workflow?
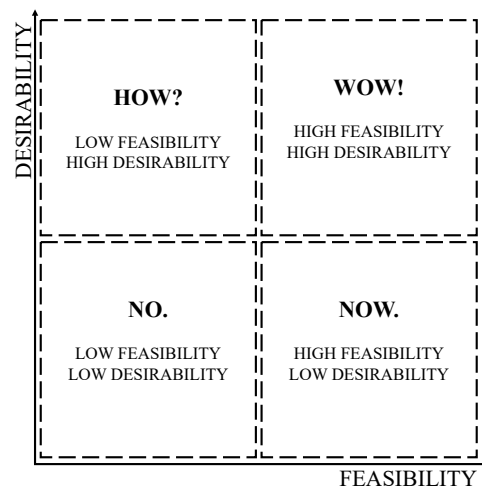
Figure 2: Feasibility-Desirability Matrix

Appendix E: Detailed Demographics

The following tables detail the expert demographics in terms of work experience, and the characteristics of the experts' organizations.

Expert	Experience (Years)	Role	Gender
ME1	2	Junior Information Security Officer	f
ME2	10	Information Security Officer	m
ME3	25	Chief Information Security Officer	m
ME4	20	Chief Information Security Officer	m
ME5	15	Chief Information Security Officer	m
ME6	35	Chief Information Security Officer	m
ME7	22	Former Chief Information Security Officer, now: Founder & Owner	m
ME8	23	Head of Security	m
ME9	15	Security Architect	m
ME10	5	Head of Security	m
OE11	18	Network Security	m
ME12	20	Chief Security Officer	m
ME13	10	Chief Security Officer	m
ME14	25	Chief Information Security Officer	m
ME15	25	Chief Information Security Officer	m
ME16	20	Information Security Officer	m
ME17	15	Chief Information Security Officer	m
ME18	14	Chief Information Security Officer	m
ME19	4	Chief Information Security Officer	m
ME20	13	Chief Information Security Officer	m
ME21	15	Chief Information Security Officer	m
ME22	12	Chief Information Security Officer and Data Protection Officer	m
ME23	20	Chief Information Security Officer	m
CE24	3	Security Consulting Engineer	f
ME25	15	Chief Information Security Officer	m
ME26	4	Chief Information Security Officer	m
ME27	10	Manager	m

Table 6: Expert description by years of experience and role

# employees	Count
<50	2
51-200	1
201-500	3
501-1000	3
1001-5000	3
5001-10000	6
>10000	6

Table 7: Experts organisations size, n=24

Industry	Count
Banking, Finance and Insurance	6
Consulting	1
Education	1
Health Services	1
IT Services	5
Manufacturing	1
Marketing	1
Media	1
Public Services	3
Telecommunication	1
Transport	2
Utilities	1

Country	Count
America	1
Germany	1
Switzerland	22

Table 8: Experts organisations industry and location, n=24

Appendix F: Autonomy Levels and Tasks

LVL	Requirements for Tasks	Task Examples
1	Personal responsibility, require contextual- or goal-understanding, or creativity	Risk management and assessment, policy development, solution architecture development
2	AI capabilities fit, non-time-critical, potentially far-reaching consequences	Patching vulnerabilities, isolating infected assets after incidents, generating content for trainings
3	Far-reaching and non-reversible consequences	Storage management
4	Time-critical and reversible, consequences could be mitigated, maintain situational awareness	Vulnerability detection, firewall configuration, monitoring network traffic, log-file analysis, detection of phishing emails
5	Routine and minimal consequences, avoid information overload	Distribute and verify user privileges, simulating phishing emails for training, malware detection

Table 9: Autonomy Levels with Tasks

Comparing Malware Evasion Theory with Practice: Results from Interviews with Expert Analysts

Miuyin Yong Wong
Georgia Institute of Technology

Matthew Landen
Georgia Institute of Technology

Frank Li
Georgia Institute of Technology

Fabian Monrose
Georgia Institute of Technology

Mustaque Ahamad
Georgia Institute of Technology

Abstract

Malware analysis is the process of identifying whether certain software is malicious and determining its capabilities. Unfortunately, malware authors have developed increasingly sophisticated ways to evade such analysis. While a significant amount of research has been aimed at countering a spectrum of evasive techniques, recent work has shown that analyzing malware that employs evasive behaviors remains a daunting challenge. To determine whether gaps exist between evasion techniques addressed by research and challenges faced by practitioners, we conduct a systematic mapping of evasion countermeasures published in research and juxtapose it with a user study on the analysis of evasive malware with 24 expert malware analysts from 15 companies as participants. More specifically, we aim to understand (i) what malware evasion techniques are being addressed by research, (ii) what are the most challenging evasion techniques malware analysts face in practice, (iii) what are common methods analysts use to counter such techniques, and (iv) whether evasion countermeasures explored by research align with challenges faced by analysts in practice. Our study shows that there are challenging evasion techniques highlighted by study participants that warrant further study by researchers. Additionally, our findings highlight the need for investigations into the barriers hindering the transition of extensively researched countermeasures into practice. Lastly, our study enhances the understanding of the limitations of current automated systems from the perspective of expert malware analysts. These contributions suggest new research directions that could help address the challenges posed by evasive malware.

1 Introduction

Malicious software or malware is a serious and constantly evolving threat to cybersecurity. As of 2023, a staggering 300,000 new malware samples are generated daily[1]. With such a large volume of new malware, it is imperative that security professionals are equipped with the tools necessary to identify and analyze these samples in a timely manner. Unfortunately, as malware becomes more sophisticated, malware authors have developed evasive techniques to make the analysis more difficult and time-consuming. In fact, in 2018 a study showed that 98% of malware samples employ at least one evasive technique[2]. A few examples of evasive techniques are code obfuscation and sandbox evasion [1, 28, 48].

Code obfuscation is a technique that deliberately makes the code more difficult to understand during static analysis, which is the process of examining a malware's functionality without executing the code. Some common examples of code obfuscation include string encryption, packing, flattening the control flow, and adding spurious code. In turn, to mitigate obfuscation techniques, researchers have developed unpackers [13, 17, 31, 49, 74, 82], and de-obfuscators [7, 18, 65, 73, 84, 94]. To impede dynamic analysis, a type of analysis that executes the malware in a controlled environment, malware authors often insert checks in their code to detect if they are being executed within an analysis environment such as a sandbox, a technique known as sandbox evasion. Such checks enable malware to evade analysis by not revealing its functionality. In response, researchers have developed more transparent sandboxes which make the analysis environment less detectable [20, 21, 26, 68, 80, 81, 87] and techniques to detect evasive samples by comparing multiple executions of the malware [30, 32, 40, 45, 53, 83]. More recently, there have been efforts such as forced execution [38, 58, 79], which aim to investigate the different execution paths of malware to expose its malicious behavior. Despite the considerable amount of research aimed at evasion techniques, a user study by Yong et al. [88] found that analyzing evasive malware continues to be a challenge for practitioners.

To keep up with the increased sophistication of evasive malware and explore how future research can enhance the analysis of evasive malware in practice, it is necessary to understand (i) methods that have been developed by past re-

---

[1]See 50+ Cybersecurity Statistics for 2023 You Need to Know

[2]See Evasive Malware Now a Commodity.

search to counter evasion, and (ii) the evasion techniques that still remain challenging for malware analysts in practice. Although prior work has conducted surveys of dynamic malware analysis evasion techniques [1, 10], none focused specifically on countermeasures that help handle evasion techniques that hinder either dynamic or static analysis. Additionally, while prior user studies [77, 88] have studied the process of reverse engineering and malware analysis, our study is the first to identify the specific evasion techniques that malware analysts in practice find challenging and examine how they currently handle such techniques. Moreover, unlike prior studies, we conduct a systematic comparison with existing literature to provide informed recommendations for future research that may help solve analysts' challenges. To meet this goal, we conduct the first systematic mapping of countermeasures for evasion techniques employed by malware, and combine it with 24 semi-structured interviews with highly experienced malware analysts who work in established security groups of well-known companies such as Proofpoint, General Electric (GE), Mandiant (now Google), IBM, and SecureWorks.

Our systematic mapping allows us to understand which malware evasion techniques have been addressed in research, and the methodologies that have been developed to counter such evasion techniques. Furthermore, our user study helps us identify evasion techniques that remain challenging in practice. Together, our systematic mapping and user study are used to perform a comparative analysis between the evasion countermeasures that research focuses on and the challenging evasive techniques practitioners encounter. This can also help inform areas in need of further research. Thus, we seek to answer the following questions:

RQ1 Which malware evasion techniques have been the focus of research dealing with evasion countermeasures?

RQ2 What are the most challenging evasive techniques encountered by malware analysts in practice?

RQ3 What approaches do malware analysts take to counter evasive techniques?

The main contributions of this paper are the following: First, we map and categorize evasive countermeasures found in the literature. Second, we identify the most challenging evasion techniques encountered by our study participants, along with the manual processes they follow to overcome such challenges. Third, we conduct a comparative analysis between solutions explored by research and challenges encountered by malware analysts in practice. Our analysis reveals that existing research solutions have significantly contributed to the field. However, there exists a misalignment between some of the practical challenges that malware experts face with evasive malware and the focus of developed research solutions. For example, we found that although malware analysts find anti-disassembly to be a significant hurdle, there is relatively less focus on research being done on anti-disassembly in the scope of malware analysis. Conversely, we found that despite

the significant amount of research on countering obfuscation techniques, participants state obfuscation as the most challenging evasion technique to handle. These observations provide valuable insights for identifying future research directions, including the development of innovative tools to assist analysts with persistent challenges and the investigation of barriers hindering the transition of existing research techniques into practice.

2 Systematic Mapping Methodology

In this section, we introduce our systematic mapping of countermeasures for evasion techniques. While there are surveys of dynamic analysis evasion techniques [1, 10], they include limited information about their countermeasures. Furthermore, to the best of our knowledge, none have covered both static and dynamic analysis evasion countermeasures. Without an understanding of previous research efforts aimed at countering malware evasion techniques, research gaps in the field remain unclear. To fill this need and provide an overview of existing research on evasion countermeasures for both static and dynamic analysis, we conduct a systematic mapping.

We chose a systematic mapping approach because it systematically identifies knowledge gaps among existing research literature and uncovers promising future research directions within the field [60, 61]. More recently, this method has gained recognition in fields such as software engineering [2, 59] and medicine [11, 62], underscoring its effectiveness in enabling a rigorous and structured overview of the current research landscape. In this study, we followed Persons's [61] guidelines, which include formulating research questions, defining the search process, establishing clear inclusion and exclusion criteria, performing data extraction aligned with the research questions, and conducting data analysis.

2.1 Mapping Research Questions

The main objective of this mapping is to identify and analyze the solutions developed in research to counter evasion techniques for malware analysis. Specifically, our mapping aims to answer the following research questions:

MQ1 Which malware analysis evasion techniques have been addressed in research?

MQ2 What methodologies are proposed by researchers to counter evasion techniques?

2.2 Search Strategy

In this study, we chose a database search as our primary search strategy. Before starting our database search, the first and second authors conducted a manual search of relevant papers to identify keywords, necessary for the creation of our search query. The manual search began by reviewing the titles of

research papers from four top security conferences (USENIX, IEEE S&P, CCS, NDSS) published between 2012 and 2022 to identify papers related to the topic of malware analysis. We scope our systematic mapping to papers addressing Windows Exe malware that employ evasion techniques because Windows is the most targeted operating system by threat actors. In fact, in 2022 Mandiant reported that 92% of the newly identified malware families run on Windows[3]. Focusing solely on Windows Exe also allows us to provide a fair comparison between our participants' challenges and research on countermeasures discussed later in §8. Subsequently, the two authors applied the following predefined inclusion criteria to the titles and abstracts:

Inclusion Criteria on Title and Abstract.

- References dynamic malware analysis, deobfuscation, unpacking, or disassembly.
- Not aimed towards mobile or IoT malware.
- Not a survey or a measurement study.

After identifying 40 papers that satisfied the above inclusion criteria, both authors read the full text of the included papers and applied the following exclusion criteria:

Exclusion Criteria on Full Text.

- The research findings does not directly help counter either static or dynamic analysis evasion techniques such as anti-sandboxing, anti-debugging, obfuscation, or anti-disassembly, nor does it provide a way for the analysis to proceed without countering the evasion techniques.

In 8 cases where the two authors disagreed, the two authors reviewed the details of the paper together to resolve disagreements, resulting in a final set of 20 papers.

To construct the database search query, we first extracted keywords from the 20 papers identified through a manual search. To extract keywords, we applied common preprocessing steps to the abstracts, including lowercasing, removing special characters, and removing stop words. While stemming and lemmatization are also common preprocessing techniques, we opted not to utilize these techniques since we require the exact words to match during search queries and these techniques would not produce exact matches. After preprocessing the data, we categorized the abstracts based on the evasion technique they address, either dynamic or static, and utilized TF-IDF analysis to identify the most significant and frequently occurring words within each category. Following this, we categorized the remaining words based on their semantic commonalities. For instance, we grouped the words debugging, automatic, and dynamic together as they all relate to the execution of a task or process. Lastly, we searched for additional synonyms used in malware research.

We used the above group of words to construct two database search queries. To make the search more precise, we decided to search only within the abstracts. We conducted the same search in IEEE and ACM databases, two of the largest research databases in computer science and engineering [4]. Two of the top security conferences, USENIX security, and NDSS, along with RAID, a conference with a historical emphasis on malware analysis, are not included in these databases. To find relevant papers published in these three conferences, we created an additional query for Google Scholar. Due to the limitations of Google Scholar search, we were not able to have an identical search query. However, the search strings were logically checked by multiple authors. All of the papers found through the database search underwent the same inclusion and exclusion criteria applied in the manual search described earlier.

2.3 Search Evaluation

To assess the quality of our search results, we conducted tests using known papers found in existing surveys [1, 10]. The first test was on 14 papers found in Table 3 from Afianian et al. [1]. While their table includes 17 papers, only 14 can be found in the databases we searched. Among these 14 papers, our initial database search was able to find 10, achieving a 71.4% retrieval rate. To improve the search, we identified reasons for the missing papers and added synonyms to the query search (ex: "avoid detection", "running", "transparently", "stealthily"). This change increased our retrieval rate to 85.7%, which is above the suggested range of 70%-80% by Kitchenham et al. [37], and added 76 papers to our total database results. To further assess the quality of our search, we conducted another test using 17 papers found in Table 4 of Bulazel et al. [10] (excluding Android and Web papers). Our database search successfully identified 14 of the 17 papers, achieving an 82.4% retrieval rate.

We narrowed our scope to papers published in class A or A* conferences (based on CORE ranking) to help ensure the quality of the papers in this study, as done in prior work [5, 55]. These conferences are known for their rigorous reviewing process. We recognize that limiting our mapping to these conferences may miss relevant papers. However, like other systematic mappings, we do not claim completeness [36, 61] though we strive to ensure quality through our evaluation.

2.4 Data Extraction and Classification

Based on our systematic mapping research questions (MQ1 and MQ2), we developed a standardized data extraction form to ensure consistency in the information gathered from each paper. This form asks for the type of evasion technique that each paper helps address, the methodology used, and the main research question being answered. The first two authors applied this form in a similar manner to reduce bias. Provided

---

[3]See M-Trends 2023

[4]We observed that the ACM database search result often included research papers that do not match the provided search query, resulting in a high number of excluded papers.
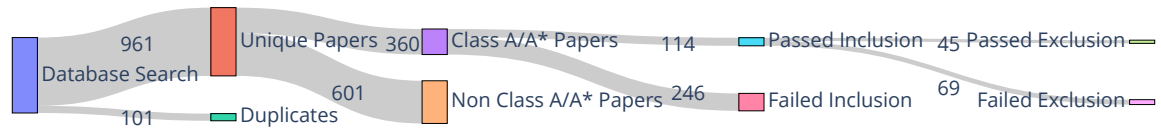
Figure 1: Papers Selection Process

Database	Dynamic Analysis Evasion Papers	Static Analysis Evasion Papers	Total Papers
IEEE	124	195	319
ACM	335	242	577
Google Scholar	86	80	166
Total Papers	545	517	**1062**

Table 1: Database Search Results

with the extracted data, the two authors were able to identify research papers that aimed to answer similar research questions and also address similar evasion techniques. Within the research papers with similar evasion techniques, we further analyzed each research paper to find commonalities among their methodologies. Through this process, we identified several papers that addressed both anti-sandbox and anti-debugging techniques so we categorized these papers into both.

3   Systematic Mapping Results

In this section, we present the results of the systematic mapping. As shown in Figure 1, with the database search, we found 1062 papers. The number of papers identified by each database query search can also be seen in Table 1. After removing duplicates and refining our search to include only class A or A*, we were left with a set of 360 papers. These 360 were then subjected to our inclusion criteria, defined in §2.2, resulting in the selection of 114 papers. Finally, after a thorough assessment of the full text, we applied our exclusion criteria and identified that the majority of the 114 papers do not directly help counter any evasion technique, which resulted in a final set of 45 papers.

Based on the number of papers that address each type of evasion technique, we find that obfuscation and anti-sandbox were the most researched evasion techniques and anti-disassembly was the least. Although we acknowledge that paper counts do not provide a complete explanation for the observed patterns, they serve as a practical and widely accepted metric for identifying trends and patterns in existing literature [2, 11, 59, 62]. Below we explain each of the four categories discovered in our mapping and provide a description of the different methodologies proposed in research to counter them. The categorized papers can be found in Table 2.

Obfuscation. Obfuscation is an evasive technique that modifies the original malware code to obscure the understanding of its functionality. Through our systematic mapping, we iden-

tified 19 papers that proposed countermeasures to this type of evasion. The majority of papers used some form of dynamic analysis to overcome the obfuscation [12, 13, 17, 49, 84, 91]. One paper by Coogan et al. [17] deobfuscates virtualized malware by identifying the system calls made when the malware executes and extracting a subtrace containing only the code related to those calls and then approximating the original code with this information. Another frequently used methodology is symbolic execution, which builds expressions containing different inputs and uses a SAT solver to find values that satisfy the expressions [7, 52, 82, 85]. One paper proposed backward-bounded dynamic symbolic execution [7], which leverages symbolic execution to answer infeasibility questions that are frequent with obfuscated code. Researchers have also applied static analysis to overcome obfuscation [46, 66]. Lu et al. [46] proposed a method to remove return-oriented programming (ROP) from malware to enable standard analysis tools to work properly on such malware.

Some papers show that a combination of static and dynamic analysis can also be effective for overcoming obfuscation [14, 63, 64]. PolyUnpack [64] is one such paper that does static analysis to build a static code model, then, during dynamic analysis, compares executed code to this model to identify when unseen code is found to unpack a sample. Finally, there are miscellaneous methodologies that are part of our mapping including program synthesis [8, 29], artificial intelligence-based search [51], and other [44].

Anti-sandbox. Anti-sandbox is an evasive technique used by malware to detect whether its execution is monitored in a controlled environment such as a sandbox. A few of the most common examples of anti-sandbox are system checks, user activity, and delay execution. Through our systematic mapping results, we identified 19 papers that propose countermeasures to this type of evasion. There are three main methodologies that these papers follow. The first one is hypervisor-based analysis [20, 43, 54]. This approach involves the use of virtualization to create isolated environments. To further improve hypervisor-based analysis, these papers focus on creating more transparent hypervisors. For example, Ether [20] proposed a novel and more transparent application of hardware virtualization extensions where the analysis engine resides completely outside the target OS environment. The second methodology is forced execution-based analysis [38, 58, 89]. Forced execution, similar to other path exploration approaches, forces malware to execute through many dif-

Evasion Types	Evasive Tactics	Methodologies	Research Papers
Static	Obfuscation	Dynamic Analysis	2007: [49], 2011: [17], 2013: [91], 2015: [84], 2018: [13], 2023: [12]
		Static Analysis	2011: [46], 2021: [66]
		Dynamic & Static Analysis	2006: [64], 2010: [63], 2021: [14]
		Symbolic Execution	2015: [52] [85], 2017: [7], 2018: [82]
		Other	2010: [29], 2017: [8], 2021: [51], 2022: [44]
	Anti-Disassembly	Dynamic & Static Analysis	2015: [9]
		Static Analysis	2004: [39]
Dynamic	Anti-Sandbox	Bare Metal-Based Analysis	2013: [93]
		Hypervisor-Based Analysis	2008: [20], 2009: [54], 2014: [43], 2015: [90]
		Introspection-Based Analysis	2016: [68], 2021: [71]
		Forced Execution-Based Analysis	2011: [38], 2014: [58], 2020: [89]
		Other	2006: [75], 2010 [16] 2011: [69], 2012: [87], 2013: [34], 2014: [83]
	Anti-Debugging	Hypervisor-Based Analysis	2013: [19], 2015: [90], 2022: [33]
		Bare Metal-Based Analysis	2015: [92]
		Introspection-Based Analysis	2016: [42] [68]
		Instrumentation-Based Analysis	2021: [26]

Table 2: Categorization of Evasion Countermeasures Research Identified Through the Systematic Mapping

ferent paths to collect additional information regarding the malware's behavior. X-force [58] specifically explores different paths without requiring specific inputs or environmental setups. They achieve this by forcing specific instructions, such as predicates and jump table accesses, to have predefined values. A third somewhat less used methodology by research is introspection-based analysis [68, 71]. Introspection refers to software's ability to examine its internal state during execution. Su et al. [71] introduces a novel Out-of-VM introspection technique called Catcher that traces malicious behavior without altering the target environment through the use of CPU cache. Finally, we found six other research papers [16, 34, 69, 75, 83, 87], each with their own distinct methodology including bare-metal-based analysis [93], static analysis [16], taint analysis [34], and multi-system execution [83].

Anti-debugging. Anti-debugging evasion techniques try to detect and prevent analysis of their code during execution. Unlike sandbox evasion techniques, anti-debugging techniques are less concerned with the execution environment and more focused on preventing the analysis of their code. However, despite these differences, they do share similar methods for countering both evasion techniques. Our systematic mapping identified 7 papers that propose countermeasures for this type of evasion. The two most common methodologies we identified were hypervisor-based analysis [19, 33] and introspection-based analysis [42, 68]. The most recent hypervisor-based analysis research paper found in our systematic mapping introduced HyperDbg [33], a specialized hypervisor-assisted debugger that relies on hardware capabilities like Intel-VT to achieve more transparency in their analysis. Conversely,

LO-PHI [68] is an example of how introspection-based analysis can be used to help counter anti-debugging techniques. LO-PHI, introduces physical hardware sensors capable of capturing memory and disk activity during execution, which can be used for analyzing evasive malware samples.

The two other methodologies in our mapping were bare metal-based analysis and instrumentation-based analysis. By executing the malware on bare metal and leveraging System Management Mode, MALT [92] enhances the transparency of the execution environment and minimizes the artifacts exposed to malware, which aids in the debugging of evasive malware. In contrast, Hong et al. [26] provide transparency for native read, write, or access to the target through a novel approach called Execution Flow Instrumentation (EFI). EFI allows a user-space program to instrument the execution flow of malicious threads across user and kernel space which can help address existing instrumentation limitations in the analysis of malware with anti-debugging techniques.

Anti-disassembly. Anti-disassembly evasion techniques format malware code in such a way that the disassembler incorrectly interprets the bytes and produces assembly code with errors. This causes problems for analysts because it prevents them from performing accurate static analysis of the malware, which is a critical component of some analysts' processes. The first work that addresses this form of evasion is by Kruegel et al. [39]. This paper primarily uses static analysis to perform a modified recursive and statistical disassembly to correctly disassemble malware that is affected by obfuscation. Bonfante et al. [9] focus on disassembling malware with self-modifying code and overlapping instructions. Their solu-

tion uses both static and dynamic analysis by first executing the malware and taking memory snapshots at different points to capture different waves of code. For each wave, the tool disassembled the code using the dynamic trace as a guide.

4   User Study Methodology

Through our systematic mapping, we were able to identify malware evasion countermeasures and the degree to which they have been explored in research. However, to assess whether these countermeasures meet practitioner needs for analyzing evasive malware, we conducted the first user study on malware evasion to determine which evasion techniques still pose challenges for analysts in practice through 24 semi-structured interviews with malware analysis experts. While an observational study was considered for our methodology, they are less common in related literature because their time-consuming nature can deter participation, and the participants' behavior may be altered during observation. For these reasons, along with NSF's guidance on qualitative methods [56], we chose to conduct semi-structured interviews for our study. Additionally, exploratory qualitative research such as semi-structured interviews offered us the flexibility to gain valuable insights into the participants' decision-making process and their challenges based on their first-hand experiences, while still providing a framework to ensure key topics were covered. Our Institutional Review Board (IRB) approved this study and participants signed a consent form before taking part in the study. To ensure the confidentiality of the participants' information, we sent a draft to all the participants prior to submission and provided an opportunity for them to review and request changes.

4.1   Recruitment

To recruit an expert group of malware analysts for this study, we utilized five different sources. We first reached out to a known Slack channel for malware analysis research and sent a description of our study to a security organization mailing list. Additionally, we reached out to personal contacts who are in the field of malware analysis. Finally, we posted a description of the study on Twitter and LinkedIn, which are not restricted to malware analysts. We specifically selected Twitter (now X) given its popularity within the security community. The description in the message explained that we were looking for malware analysts to participate in a research study in order to understand the current state of practice of evasive malware analysis. We reached out to all five sources listed above in November 2022, and included a link to the initial questionnaire for this study.

Participant Screening. Given the wide range of recruitment sources, we had an initial participant screening phase. To verify that the participants have experience with evasive malware, we had prospective participants fill out a questionnaire with

22 questions hosted on Microsoft Forms. We estimated this questionnaire would take 10 minutes to complete. The first set of questions was about their background information and analysis objectives. The majority of the questions focused on their experience analyzing evasive malware. The last subset of questions contained optional, demographic questions. After interviewing our first participant and understanding the significance of anti-disassembly techniques, we decided to add the question "Based on your experience, which of the following categories of anti-analysis is the most challenging?" This decision was made after meticulously considering the number of responses received and determining a way to ask this question to the participants who had already submitted the questionnaire. Given that we only received 7 responses at the time, we decided to modify the questionnaire for future responses and asked this question during the interview to the participants who had previously submitted the questionnaire. The complete questionnaire is provided in Appendix A.

We received a total of 109 responses to the questionnaire. We analyzed each response by determining whether the answers made sense in the context. Additionally, we attempted to verify the identity of those who responded by looking at the requested LinkedIn profiles. 75 responses were discarded due to random responses to questions that did not indicate the responder had malware analysis skills. Additionally, we were not able to confirm the identity of 5 responses through LinkedIn, so we reached out to them and requested additional information that could help us confirm their identity. Unfortunately, they did not respond, so we were forced to exclude them. We invited the remaining 27 respondents to participate in the interview process and 24 scheduled an interview.

4.2   Interview Protocol

For each participant, we conducted an hour-long, semi-structured interview via online video call. For consistency, all interviews were conducted by the same researcher from November 2022 through January 2023. The interview was broken down into three sections, described below, totaling 29 questions, as seen in Appendix B.

Identifying and Analyzing Evasive Malware. The interview began with understanding how participants identify that a malware sample is evasive and how their analysis process differs when analyzing an evasive sample. Additionally, we asked the participants to explain the main challenges they encounter when analyzing evasive malware samples.

Techniques Used for Handling Evasive Malware. The second section of questions focuses on understanding how malware analysis practitioners handle different evasion techniques. We asked them to walk us through examples of how they handle different evasive techniques such as sandbox evasion, and obfuscation. Additionally, we asked the participants what are the most time-consuming and challenging evasive techniques they encounter, and how they handle them.

Use of Existing Tools for Malware Analysis. The last part of the interview questions is about the analysis tools that the participants use during their analysis process. We asked them what tools they use and which are the most helpful. We also wanted to know what aspects of the workflow would benefit from a new automated system and what improvements they want to see. The purpose of these questions was to understand challenges that need to be addressed.

To ensure that the interview questions were complete, we conducted 3 pilot interviews with graduate students who have malware analysis experience and incorporated their feedback.

4.3 Data Collection and Analysis

All the interviews were audio recorded and automatically transcribed with the built-in video call software. To make sure the transcriptions were accurate, the two first authors manually reviewed and corrected transcription errors. The same researchers then coded the interviews using an iterative open coding methodology [70]. First, they independently coded the first 4 interviews and then agreed on an initial set of codes. These codes were then used by each researcher to recode the same 4 interviews. After this process, we compared the codings between the researchers and eliminated codes that were either redundant or too specific and developed the final codebook composed of 80 codes, as seen in Appendix C. With this final codebook, the authors coded the remaining 20 interviews. The Krippendorff's alpha intercoder reliability score was 0.95 [41], indicating very high consistency. The codes from the final codebook are used to identify patterns and produce our results in §6 and §7.

To determine the adequate amount of participants required for our study, we calculated the point at which we reached saturation. Saturation occurs when no new novel themes are found with additional interviews. To compute saturation, we followed Guest et al. [23] by looking at the interviews that were coded after the final codebook was developed and determining how many interviews it took for all the codes to appear. Additionally, we confirmed that no novel themes emerged. Following this procedure, we reached saturation after the 9th interview. To further validate our findings, when we shared a draft of the paper with the participants, we received several responses stating that they enjoyed reading how their personal experiences aligned with our findings.

5 User Study Participants

We interviewed a group of 24 malware analysts. Most participants have more than 7 years of experience with malware analysis and thus can be viewed as expert analysts. By interviewing participants with many years of experience working at well-established security groups in companies such as Mandiant (now Google), GE, Proofpoint, and Cisco, we were able to identify the most challenging evasive techniques and obtain a broad understanding of how evasive malware is handled in practice. By far, the most common degrees completed by the participants were computer science and electrical engineering. Some participants also specialized in related fields such as information technology or computer networks. Finally, 10 participants stated that they acquired their malware analysis skills through a mentor who played a crucial role in shaping their current analysis workflow. A detailed table of our participants' backgrounds can be found in Table 3 in Appendix D.

6 Malware Analysts' Perspective

In this section, we present malware analysts' definition of evasion and their most challenging evasion techniques.

6.1 Definition of Evasion

To gain a better understanding of what practitioners consider evasive malware, we asked each participant to define evasive malware. The majority of the participants consider evasion to be any technique that affects either their static or dynamic analysis process. As P10 said, an evasive sample "has code, which has the primary goal, intended or not, of disrupting analysis. There are a couple of different forms of analysis, dynamic and static, and each of them has [evasion techniques] to make analysis harder." Static analysis evasion techniques are "designed to make static analysis difficult, such as through obfuscation of code and/or data, or other techniques that incur additional steps to deobfuscate it," as P20 states. Another static technique is anti-disassembly, which causes the disassembler to recover incorrect instructions from the binary. To evade dynamic analysis, malware authors "either bypass the instrumentation itself, detect the environment, or logically not expose behaviors. [This can be done through] delayed execution [or by] requiring parameters," as P7 expressed. Dynamic analysis evasion can affect the proper execution of the malware sample in a debugger, sandbox, or any other controlled environment analysts may use to analyze the sample.

6.2 Most Challenging Evasive Techniques

One of the main objectives of this study is to determine the most challenging evasion techniques malware analysts encounter. To answer this question (RQ2), in the questionnaire, we asked what is the most challenging anti-analysis technique they routinely need to handle. The provided options were anti-disassembly, anti-debugging, sandbox evasion, and others. We further refine the option anti-disassembly into 2 categories; obfuscation, and anti-disassembly based on participants' explanations during the interview. This change also facilitates the comparison between malware analysts' challenges and research countermeasures discussed later in §8.

According to the responses provided by the study participants obfuscation was identified as the most challenging

evasion technique analysts routinely have to handle by 9 participants. As P1 explained, "I think in terms of what is probably the biggest problem on the team, just across all malware right now, it's control flow obfuscated malware. That obfuscation can show up in any kind of malware. JavaScript, PowerShell, windows PEs, you name it. It'll be everywhere." Control flow obfuscated malware tries to hide the actual flow of instructions, which makes it hard for analysts to understand the malware's logic. The second most challenging evasion technique was anti-disassembly, which was mentioned by 6 participants. Anti-disassembly techniques affect static analysis by hindering the proper function of the disassembler. P11 explains that "it tends to be very hard to automate your way out of. It's hard to make a generic anti-anti-disassembly tool."

3 participants expressed anti-debugging as being the most challenging. Anti-debugging techniques include checking for the presence of a debugger or altering the code to prevent reverse engineers from stepping through the code while running it in a debugger. As P20 explains, this evasive technique is challenging because "if it is designed in a way that I can't even follow the code execution [...] that makes it really difficult to figure out which blocks I should narrow in on for doing static analysis work, and it makes it really difficult to create detection signatures." Furthermore, only 2 participants mentioned sandbox evasion to be the most challenging. The reason why sandbox evasion is not considered much of a challenge is "largely because the anti-sandbox stuff I could [...] just run out on a real system, and that real system is still instrumented with a lot of the same tools." as P20 stated.

7 Workflows For Handling Evasive Tactics

To inform future research on ways to address the challenges posed by evasion techniques, we studied how expert malware analysts counter evasion techniques to further analyze the malware's behavior. This information gave us an opportunity for an in-depth look into the processes that each analyst follows. It was not surprising to see that each participant employs a somewhat different workflow, as it can be argued that the process of malware analysis involves a certain degree of creativity, similar to an art form. Despite the creative and varied nature of this process, we were able to distill commonalities among participants' workflows and generated 3 distinct workflows malware analysts follow to handle evasive malware.

7.1 How Malware Experts Handle Dynamic Analysis Evasion

To handle dynamic analysis evasion, our participants use one of two workflows, alter the dynamic analysis execution, or resort to static analysis.

Debugging Workflow for Forced Execution. *Workflow 1*,
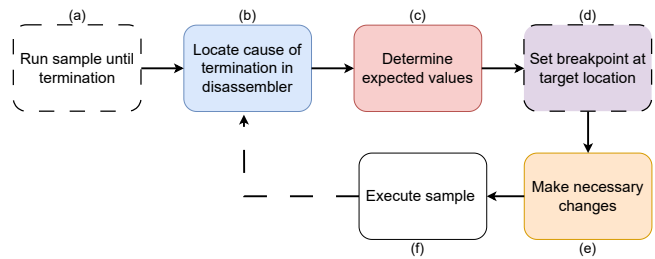
Figure 2: Workflow 1: Forced Execution; P1-P4, P6, P9-P12, P14, P17, P18*, P20* P21, P24*. The asterisk symbol highlights participants that also patch the malware

the first workflow to handle dynamic analysis evasion, shown in Figure 2, is used by 15 of the participants. It is important to note that any line or task with a dotted line is optional. This workflow is used by analysts who either make changes in real-time in the debugger to force the malware sample to execute or patch the malware to create a new version of the malware without evasive checks. To begin, participants execute the malware sample until termination (step a). This can be done in a debugger or a sandbox. Once the sample terminates, the analyst uses the end of the execution trace to locate the code responsible for the evasive check in the disassembly (step b). As P1 stated, "if I'm looking at a sample and I look at the trace and notice that it's stopped there, I can pull it up in IDA, and go right to where it happens." Alternatively, other participants begin this workflow in step b by locating the evasive check without executing it in the debugger. As P4 explained, "the most important parts to understand the [malware's] logic are the conditional branches because that's where the malware is checking, is this true? [...] Is it greater or less than?." After locating the dynamic evasion check, the analysts utilize static analysis to better understand the malware's functionality and determine why the sample stopped executing (step c). This step may also involve identifying the necessary values required for the malware to continue executing.

After gaining a better understanding of the malware's functionality and determining the expected values, analysts either move to step d or step e. The majority of the participants stated a preference towards step d, where they set a breakpoint in the debugger right before the function containing the dynamic evasion technique (step d). Once they execute the sample and reach the breakpoint, the analysts change the value of registers, alter portions of dynamic memory, or flip appropriate bits of conditional jumps to force the malware to continue executing (step e). P17 explained his process as "sometimes it's just as easy as, I change this from 0 to 1, all of a sudden I get a whole new branch. Sometimes it's more complicated [...] and then I have to go into the debugger, change a register or a portion in dynamic memory to read something else, to force the program to do what I want it to do." After the dynamic evasive check is handled, the sample will continue executing until it reaches the next point of termination (step f), at which point, the analyst may decide to repeat this workflow

until the analyst is able to handle additional evasive checks and complete the analysis.

Debugging workflow to execute target functions. One major disadvantage of *Workflow 1* is that analysts have to repeatedly apply it for every evasive check until they reach their goal, which could be tedious, time-consuming, and worse, does not scale. As P2 explains, "it might not always be feasible to get the full code to run because maybe a sample does multiple checks or it's not as straightforward as defeating a check once and letting it execute." To ease this burden, 5 participants prefer to do a more targeted analysis following *Workflow 2* shown in Figure 3. Although the first three steps of this workflow are similar to *Workflow 1*, one main difference is that analysts can choose the point where they begin executing the sample (step d). This point can be after the execution of dynamic analysis evasion techniques in order to avoid handling them. As P1 said, "then I'll just pick random spots after it to start execution again." Analysts may also decide to execute parts of the malware sample that they may find challenging to understand statically. P22 provides an example of when they have used *Workflow 2*, "sometimes it's hard to wrap your mind around an algorithm you're seeing in static analysis. A big loop or some sort of mathematical transform. [...] So you can identify it statically and then go run it." Because the analyst skipped an initial part of the execution, they may have to manually configure the memory and registers for the malware to run properly (step e). This could involve, "thinking about what is being passed to a function and you might have to sort of fabricate parameters to be passed to that code.", as P2 explained. Once the sample is properly configured, the analyst can run the target function and obtain the return value (step f).

Alter Dynamic Analysis Execution. Based on *Workflow 1* and *Workflow 2*, it is evident that the primary tool employed by malware analysts for countering dynamic analysis evasion is a debugger. As reported by 22 participants, the inclination towards using a debugger may be attributed to the limitations of dynamic analysis systems, which fail to capture the entirety of evasive malware's behavior. To mitigate this limitation and handle dynamic analysis evasion, 16 participants either execute the sample in a different system or make alterations to their dynamic analysis system of choice. In such cases, 9 participants choose to run the sample in either commercial sandboxes, internal sandboxes, or a bare metal system, which are more resilient against dynamic analysis evasion. Some of the commercial sandboxes that the participants referred to were VMray, Any.run, Hybrid Analysis, FLARE Sandbox, and Joe Sandbox. Participants stated that these sandboxes incorporate many anti-evasion techniques to handle the most known dynamic analysis evasive techniques. Some participants have access to internal sandboxes that are also able to handle dynamic analysis evasive tactics. As P17 explained, "anytime we come across something like that, like a timing check or a new technique that's trying to look at the environment, we try to build that into the sandbox so that next time
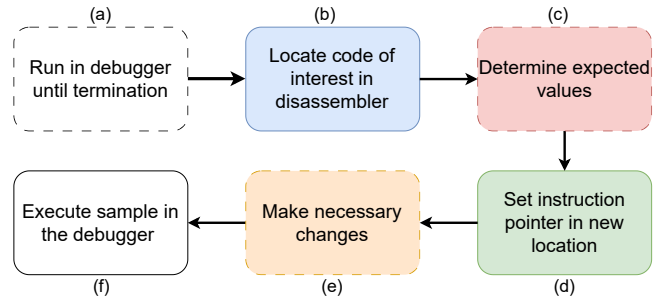
Figure 3: Workflow 2: Targeted Execution; P1, P2, P6, P9, P22

the analyst spins up that sandbox, they don't have to worry about patching over it."

Although these can be effective techniques, not all of the participants have access to such sandboxes. Another option is to manually alter the execution environment to handle the dynamic analysis evasion techniques. As P23 said, "let's try giving a different, fake username just to see [...] in the process of reversing it to try to figure out what it's doing, sometimes you take guesses and think [...] I'm going to try something instead of spending another 2 hours trying to statically reverse it. Let me just try something in 2 minutes with another dynamic analysis run that has some different option." This process generally requires static analysis to determine what changes to make although sometimes analysts will make educated guesses about what changes to make, based on experience. When the analysts are not able to quickly find necessary changes, they often utilize *Workflow 1*, *Workflow 2* or decide to statically analyze the sample.

Analyze the Sample Statically. A few participants who deal with dynamic analysis evasion rely entirely on static analysis as their primary approach. As P11 explained, "since I'm really comfortable with static analysis as opposed to dynamic analysis, I usually just blow through those anti-dynamic analysis measures pretty quickly and I'll just look at it statically and get what I need from it there." This approach is effective for samples that include dynamic analysis evasion techniques but not sophisticated static analysis evasion techniques.

7.2 How Malware Experts Handle Static Analysis Evasion

To handle static analysis evasion, our participants either use workflow 3 or use workflows 1 or 2.

Debugging Workflow for Unpacking. As shown in Figure 4, *Workflow 3* is used by 8 participants primarily to unpack a sample. Additionally, this workflow can also be used to semi-automate the de-obfuscation process. Analysts have expressed that it's easier to execute the malware to de-obfuscate itself than to go through the process statically. For example, P2 said, "I'll just find that in the debugger and let it do all the decoding and then I'll just see what it decoded." The first step in this workflow is to locate the code of interest, which in
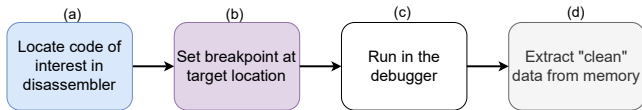
Figure 4: Workflow 3: Unpacking; P2, P4, P5, P6, P7, P20, P22

this case is the function that is responsible for unpacking or deobfuscating the code (step a). This can be done by opening up the binary in the disassembler and "identifying that that's the function that does the string deobfuscation," as P6 stated. After locating this function, the analyst launches a debugger and sets a breakpoint after the identified location (step b), and executes the sample (step c). When the execution reaches the breakpoint, the analyst extracts the deobfuscated data from memory, such as decrypted code, or plain text strings (step d).

Debugging Workflow for Targeted De-obfuscation. *Workflow 3* is an effective way for analysts to handle static analysis evasion when the sample does not include dynamic analysis evasion before the deobfuscation function. However, 10 participants have mentioned that some malware samples implement both static and dynamic analysis evasion. In such cases, the analysts rely on either *Workflow 1* or *Workflow 2*. When utilizing *Workflow 1*, the participants begin executing the sample from the entry point and continue handling each dynamic evasive check until they reach the function that deobfuscates the encrypted data. Once the deobfuscation has been completed, the analyst extracts the data from memory in the same way as the previous workflow.

Although this strategy is effective, participants have expressed that in some cases, malware samples either implement dynamic analysis evasive techniques that are more difficult to handle or include too many dynamic analysis evasive checks, which can be time-consuming. To reduce the analysis time for such samples, participants prefer to use *Workflow 2*, where the code of interest is the function that deobfuscates the target data, such as string decryption. One interesting finding that P2 mentioned is the scenario in which the target function requires parameters to be passed. P2 said "it's getting 4 parameters, what are these parameters? The first one might be where the payload is in the code, the second one might be a key, [...] you might have to do a little bit of work upfront to sort of force it to execute." Following *Workflow 1* and *Workflow 2* allows the participants to extract the information that they want without having to reimplement the samples' decryption algorithm.

8   Comparing Malware Analysts Evasion Challenges and Research Countermeasures

To compare how evasion countermeasures explored by research align with challenges faced in practice, we conducted a comparative analysis between the challenges (§6.2) highlighted by our participants and the solutions found in existing research (§3). As illustrated in Figure 5, although 25% of the participants' stated anti-disassembly as the second most
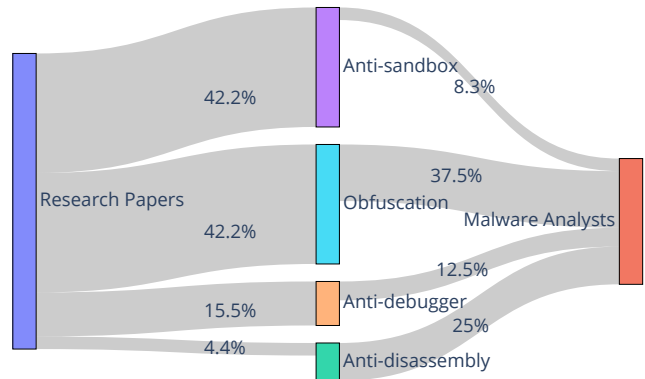
Figure 5: Research on Evasion Technique Countermeasures Vs. Challenging Evasion Techniques for Malware Analysts

challenging evasive technique to handle, based on the volume of identified papers in our systematic mapping, it appears to be less emphasized in research. Our systematic mapping found that this category represented only 4.4% of the research papers, suggesting an opportunity for future research to delve more into this challenging task. Countermeasures for anti-disassembly are especially important now given that participants report a rising trend of malware samples developed in uncommon programming languages, which may lead to inaccuracies in the disassembly and debugging process. As P2 explained, "alternative languages are becoming problematic. So like Golang, Rust, and Delphi are three languages that when you write a program and compile it, it is a lot less straightforward than looking at compiled C." In fact, some participants consider the use of such languages to be a novel evasive technique and participants report that there are not many available tools to deal with this rising problem. Without tools to accurately recover the malware's instructions when it is written in these non-standard languages, analysts' static analysis will be more challenging and less accurate.

As discussed in §6.2, obfuscation is by far the most challenging evasive technique for malware analysts to handle, with 37.5% of the participants stating this observation. At the same time, this evasive technique is by far the one with the most research aimed at developing countermeasures and accounts for 42.2% of the analyzed papers. Based on this observation, we suggest future research directions in §9.

More interestingly, sandbox evasion has the same volume of research papers (42.2%) as obfuscation but is only considered to be a significant challenge by 8.3% of the participants due to their experience with malware analysis, which allows them to circumvent sandbox evasion through static analysis or a debugger, as mentioned in §7. It is worth noting that research has made a lot of strides in creating more resilient and stealthy sandboxes to help handle many techniques used for fingerprinting or environment detection [20, 35, 75, 81, 87]. Despite analysts' ability to handle sandbox evasion techniques, most participants expressed a desire for additional

capabilities. For example, P7 stated that "pure dynamic analysis is very effective for the first stage [...] However, in the second, third to fourth stages, you at least need to somehow convince the attacker to send you them. It might happen not in 4 min. It might happen in 4 days." In such cases, the sandbox may not be able to trigger the behavior that executes each stage of the malware. Without all the stages, analysts are not able to complete their analysis.

Lastly, anti-debugging techniques were identified to be the most challenging evasion to handle by 12.5% of our study participants and account for 15.5% of the papers found through our systematic mapping. Based on the participants' statements, analysts are not in need of additional tools to handle anti-debugging because available tools such as Scyllahide that can be used to handle the majority of these techniques.

9 Discussion

Through our comparative analysis in §8, we are able to obtain a better understanding of the impact of research on the state of practice of evasive malware analysis. Specifically, we identified discrepancies between research and practice, finding instances where challenging techniques lack sufficient research attention and others where significant research exists despite persistent analyst difficulties. To help mitigate the identified discrepancies between research and practice, we discuss future research directions that could help analysts address the challenges they still face when analyzing evasive malware. Additionally, we underscore the need to investigate barriers that impede the transition of research into practice.

Analysis of Malware with Anti-disassembly Techniques. Our results show that anti-disassembly is perceived as a major challenge by many participants, yet in our systematic mapping we found a noticeable gap in research focused on countering evasive malware that implements these techniques. While we acknowledge the difficulties in addressing the challenges of anti-disassembly, our observations reveal a prevalent reliance on disassemblers among malware analysts in practice. Consequently, it is imperative for future research to develop techniques specifically tailored for analyzing malware with anti-disassembly capabilities. More specifically, study participants highlighted a need for ways to counter such evasive malware written in less common programming languages such as Golang, as mentioned in §8. Effective methodologies that counter anti-disassembly can significantly reduce the amount of effort required by analysts to locate the cause of termination (step b in *Workflow 1*), locate the code of interest (step b in *Workflow 2*, step a in *Workflow 3*), and understand the malware's logic when it is written in alternative languages or has evasive techniques that affect the disassembly process.

Evaluation of Existing Research Solutions in Practice (De-obfuscation Tools). Obfuscation is another serious challenge that was mentioned by our study participants. However, considering that previous research has developed many automated systems to deobfuscate malware, as shown in §3, it raises the question of why obfuscation remains a major challenge in analyzing evasive malware in practice. Future research should focus on identifying and addressing any potential barriers that impede the transfer of this research into practice.

Malware Analysis Research with Analysts in Mind. In our analysis, we identified the need for sandboxes tailored to meet malware analysts' specific needs. Prior work on sandboxes focuses on scalability and tries to optimize execution time [40]. As expert analysts focus on more sophisticated malware that employs evasive techniques, they require a more granular report with a longer execution time. As P17 said, "I would rather wait a longer time for a tool to go through and really explore an executable, and automatically go through all the different branches that could be taken." One promising direction for meeting this requirement is symbolic execution. Contrary to forced execution solutions [38, 58, 89], symbolic execution maintains a valid execution state of the malware at all times and is less likely to miss unexpected execution flows that could reveal the malware's malicious behavior. This approach is further motivated by 4 participants who mentioned they use angr [67], a popular open-source binary analysis framework that applies symbolic execution. Unfortunately, deploying this tool in analysts' workflow is difficult because it's not targeted for malware analysis. As P10 said, "angr has very limited applications in malware analysis, and things like a debugger or emulation will work a lot better for me".

To design tools that better meet malware analysts' needs and that more easily integrate with analysts' current analysis process, researchers could follow *Workflow 1* as a guide. A tool could begin by executing the malware sample until it terminates (step a), then it could automatically locate the last conditional branch that was executed before termination through static analysis (step b). The next step in the analysis workflow could be to determine what inputs the malware requires to satisfy the conditional branch (step c). This step is currently a demanding manual process that could benefit from the implementation of symbolic execution. By providing the symbolic execution engine with the conditional branch as a target, it could come up with constraints for the path that did not terminate and generate new inputs that allow the malware to continue executing. Once the inputs are generated, they can either be provided to the malware analyst or set automatically in the dynamic analysis environment to continue executing the malware sample. The implementation of such systems could help analysts observe more of the malware's behavior in a shorter amount of time. Additionally, these systems could also help analyze malware samples that implement both static and dynamic analysis evasion, as participants mentioned the use of Workflow 1 as a way of analyzing these malware samples in §7.2. Ideas along these lines appear in the work of Chipounov et al. [15] but they are applied in the context of reverse engineering and bug finding. Based on our findings, we

believe that these recommendations can help narrow the gap between research advances and practitioner needs, thereby enhancing the overall impact of research in the field.

10   Limitations

Exploratory qualitative research practices have well-known limitations. The first limitation is the lack of complete recall of participants' analysis process [25]. To mitigate this limitation, we asked participants to state as much as they could recall and only after they were done, move to the next question. This is a best practice used to ensure recall [27]. The second limitation is that participants may modify their answers to appear more experienced in the field. We aimed to mitigate this limitation by confirming their identity and years of experience through the participant screening described in §4.1. Given the exploratory nature of this study, the number that we provide in each finding is the number of participants that explicitly stated a concept. However, we acknowledge that this number may be higher as some participants may have failed to state the concept.

As was mentioned previously in §2, we scoped our systematic mapping results to papers published at tier A or tier A* conferences which may exclude some relevant papers. However, in line with the nature of systematic mappings, we do not claim completeness [36, 61]. Instead, we focus on ensuring the quality of our mapping through a rigorous evaluation. It is also important to recognize that although paper counts alone do not offer a comprehensive explanation for the observed patterns, they serve as a practical and widely accepted metric for identifying patterns within existing literature [2, 11, 59].

11   Related Work

To our knowledge, only four human-centered studies have been conducted to understand the cognitive process of software reverse engineering [4, 47, 77, 88]. Of those, two are focused on malware analysis [4, 88]. The first user study with reverse engineers as participants was conducted by Votipka et al. [77] in 2020. This study used semi-structured observational interviews to gain an understanding of the process of reverse engineering. The authors were able to develop workflows that represent the necessary process reverse engineers follow and suggest guidelines for designing future reverse engineering tools. This early research has informed subsequent studies that further explore this area [47], and investigate other related fields such as malware analysis [88]. In 2021, Yong et al. [88] conducted a user study specifically to understand the objectives and workflows of malware analysts in practice. In this study, they proposed a taxonomy of malware analysts and identified workflows employed by the participants. In contrast to [88], we conduct a systematic mapping of countermeasures for malware evasion techniques, which allows us to understand the extent to which malware evasion techniques have been researched. Additionally, we perform a user study to identify the specific evasion techniques that remain challenging for malware analysts in practice. More importantly, our work identifies potential gaps by comparing evasion techniques that remain challenging for practitioners with countermeasures explored in the research literature, which is not addressed in [88]. Thus, other than a similar user study methodology, there is minimal overlap between our research.

Montovani et al. [47] aim to understand the mental model and strategies adopted by reverse engineers to solve static RE tasks. Unlike previous studies using interviews, Montovani et al. designed a web-based platform similar to traditional interactive disassemblers, which allowed a fine-grained observation of the participants' actions. Their findings revealed expert REs visit fewer basic blocks than beginner REs, and identified strategies strongly correlated with experience level.

Recent work by Aonzo et al. [4] compares the procedures followed by humans and machines to classify unknown programs as benign or malicious, aiming to understand how data from malware analysis reports is used to reach a decision. They accomplish this by designing an online game that requests participants to classify suspicious files based on their sandbox reports. During this activity, features required by the analysts are observed. These features are then compared to two Machine Learning (ML)-based malware classification models. The key finding of this study is that the ML algorithms performed less effectively and do not use the same features as the malware analysts. ML focuses on static features, while humans rely more on dynamic features.

Although it is only recently that researchers have started to study cognitive procedures in malware analysis and reverse engineering, it is important to acknowledge the considerable amount of usable security research. Its predominant focus has been on secure software development [6, 22, 57, 72, 76] and vulnerability detection [3, 24, 78]. Prior work has also tested the usability of tools used by reverse engineers [50, 86].

12   Conclusion

In this study, we focus on malware evasion techniques and their countermeasures. We conduct a systematic mapping of research in evasion countermeasures. Additionally, we conducted a user study with malware experts to determine the most challenging evasive techniques that analysts encounter in practice. This allowed us to identify three distinct workflows that capture the process that analysts use to handle malware evasion. Lastly, we performed a comparative analysis of solutions explored by research and challenges encountered in practice. Such analysis can inform future research directions that can help analysts handle challenging evasive techniques. It can also help understand potential barriers that may be hindering the transition of research into practice.

References

[1] Amir Afianian, Salman Niksefat, Babak Sadeghiyan, and David Baptiste. "Malware dynamic analysis evasion techniques: A survey". In: *ACM Computing Surveys (CSUR)* 52.6 (2019), pp. 1–28.

[2] Muhammad Ovais Ahmad, Denis Dennehy, Kieran Conboy, and Markku Oivo. "Kanban in software engineering: A systematic mapping study". In: *Journal of Systems and Software* 137 (2018), pp. 96–113.

[3] Noura Alomar, Primal Wijesekera, Edward Qiu, and Serge Egelman. "" You've got your nice list of bugs, now what?" vulnerability discovery and management processes in the wild". In: *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*. 2020, pp. 319–339.

[4] Simone Aonzo, Yufei Han, Alessandro Mantovani, and Davide Balzarotti. "Humans vs. Machines in Malware Classification". In: *32nd USENIX Security Symposium (USENIX Security 23)*. 2023, pp. 1145–1162.

[5] Daniel Arp, Erwin Quiring, Feargus Pendlebury, Alexander Warnecke, Fabio Pierazzi, Christian Wressnegger, Lorenzo Cavallaro, and Konrad Rieck. "Dos and don'ts of machine learning in computer security". In: *31st USENIX Security Symposium (USENIX Security 22)*. 2022, pp. 3971–3988.

[6] Hala Assal and Sonia Chiasson. "Security in the Software Development Lifecycle." In: *Fourteenth symposium on usable privacy and security (SOUPS 2018)*. 2018, pp. 281–296.

[7] Sébastien Bardin, Robin David, and Jean-Yves Marion. "Backward-bounded DSE: targeting infeasibility questions on obfuscated codes". In: *2017 IEEE Symposium on Security and Privacy*. IEEE. 2017, pp. 633–651.

[8] Tim Blazytko, Moritz Contag, Cornelius Aschermann, and Thorsten Holz. "Syntia: Synthesizing the Semantics of Obfuscated Code." In: *26th USENIX Security Symposium (USENIX Security 17)*. 2017, pp. 643–659.

[9] Guillaume Bonfante, Jose Fernandez, Jean-Yves Marion, Benjamin Rouxel, Fabrice Sabatier, and Aurélien Thierry. "Codisasm: Medium scale concatic disassembly of self-modifying binaries with overlapping instructions". In: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. 2015, pp. 745–756.

[10] Alexei Bulazel and Bülent Yener. "A survey on automated dynamic malware analysis evasion and counter-evasion: Pc, mobile, and web". In: *Proceedings of the 1st Reversing and Offensive-oriented Trends Symposium*. 2017, pp. 1–21.

[11] Anna Cantrell, Elizabeth Croot, Maxine Johnson, Ruth Wong, Duncan Chambers, Susan K Baxter, and Andrew Booth. "Access to primary and community health-care services for people 16 years and over with intellectual disabilities: a mapping and targeted systematic review". In: (2020).

[12] Binlin Cheng, Erika A Leal, Haotian Zhang, and Jiang Ming. "On the Feasibility of Malware Unpacking via Hardware-assisted Loop Profiling". In: *32nd USENIX Security Symposium (USENIX Security 23)*. 2023, pp. 7481–7498.

[13] Binlin Cheng, Jiang Ming, Jianmin Fu, Guojun Peng, Ting Chen, Xiaosong Zhang, and Jean-Yves Marion. "Towards paving the way for large-scale windows malware analysis: Generic binary unpacking with orders-of-magnitude performance boost". In: *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. 2018, pp. 395–411.

[14] Binlin Cheng, Jiang Ming, Erika A Leal, Haotian Zhang, Jianming Fu, Guojun Peng, and Jean-Yves Marion. "{Obfuscation-Resilient} Executable Payload Extraction From Packed Malware". In: *30th USENIX Security Symposium (USENIX Security 21)*. 2021, pp. 3451–3468.

[15] Vitaly Chipounov, Volodymyr Kuznetsov, and George Candea. "S2E: A platform for in-vivo multi-path analysis of software systems". In: *Acm Sigplan Notices* 46.3 (2011), pp. 265–278.

[16] Paolo Milani Comparetti, Guido Salvaneschi, Engin Kirda, Clemens Kolbitsch, Christopher Kruegel, and Stefano Zanero. "Identifying dormant functionality in malware programs". In: *2010 IEEE Symposium on Security and Privacy*. IEEE. 2010, pp. 61–76.

[17] Kevin Coogan, Gen Lu, and Saumya Debray. "De-obfuscation of virtualization-obfuscated software: a semantics-based approach". In: *Proceedings of the 18th ACM conference on Computer and communications security*. 2011, pp. 275–284.

[18] Mila Dalla Preda, Matias Madou, Koen De Bosschere, and Roberto Giacobazzi. "Opaque predicates detection by abstract interpretation". In: *Algebraic Methodology and Software Technology: 11th International Conference, AMAST 2006, Kuressaare, Estonia, July 5-8, 2006. Proceedings 11*. Springer. 2006, pp. 81–95.

[19] Zhui Deng, Xiangyu Zhang, and Dongyan Xu. "Spider: Stealthy binary program instrumentation and debugging via hardware virtualization". In: *Proceedings of the 29th Annual Computer Security Applications Conference*. 2013, pp. 289–298.

[20] Artem Dinaburg, Paul Royal, Monirul Sharif, and Wenke Lee. "Ether: malware analysis via hardware virtualization extensions". In: *Proceedings of the 15th ACM conference on Computer and communications security*. 2008, pp. 51–62.

[21] Daniele Cono D'Elia, Emilio Coppa, Federico Palmaro, and Lorenzo Cavallaro. "On the dissection of evasive malware". In: *IEEE Transactions on Information Forensics and Security* 15 (2020), pp. 2750–2765.

[22] Kelsey R Fulton, Daniel Votipka, Desiree Abrokwa, Michelle L Mazurek, Michael Hicks, and James Parker. "Understanding the how and the why: Exploring secure development practices through a course competition". In: *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*. 2022, pp. 1141–1155.

[23] Greg Guest, Emily Namey, and Mario Chen. "A simple method to assess and report thematic saturation in qualitative research". In: *PloS one* 15.5 (2020), e0232076.

[24] Marco Gutfleisch, Jan H Klemmer, Niklas Busch, Yasemin Acar, M Angela Sasse, and Sascha Fahl. "How does usable security (not) end up in software products? results from a qualitative interview study". In: *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2022, pp. 893–910.

[25] Erik Hollnagel. *Handbook of cognitive task design*. CRC Press, 2003.

[26] Jiaqi Hong and Xuhua Ding. "A novel dynamic analysis infrastructure to instrument untrusted execution flow across user-kernel spaces". In: *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2021, pp. 1902–1918.

[27] Stacy A Jacob and S Paige Furgerson. "Writing interview protocols and conducting interviews: Tips for students new to the field of qualitative research." In: *Qualitative Report* 17 (2012), p. 6.

[28] Ashish Jadhav, Deepti Vidyarthi, and M Hemavathy. "Evolution of evasive malwares: A survey". In: *2016 International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT)*. IEEE. 2016, pp. 641–646.

[29] Susmit Jha, Sumit Gulwani, Sanjit A Seshia, and Ashish Tiwari. "Oracle-guided component-based program synthesis". In: *Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering-Volume 1*. 2010, pp. 215–224.

[30] Noah M Johnson, Juan Caballero, Kevin Zhijie Chen, Stephen McCamant, Pongsin Poosankam, Daniel Reynaud, and Dawn Song. "Differential slicing: Identifying causal execution differences for security applications". In: *2011 IEEE Symposium on Security and Privacy*. IEEE. 2011, pp. 347–362.

[31] Min Gyung Kang, Pongsin Poosankam, and Heng Yin. "Renovo: A hidden code extractor for packed executables". In: *Proceedings of the 2007 ACM workshop on Recurring malcode*. 2007, pp. 46–53.

[32] Min Gyung Kang, Heng Yin, Steve Hanna, Stephen McCamant, and Dawn Song. "Emulating emulation-resistant malware". In: *Proceedings of the 1st ACM workshop on Virtual machine security*. 2009, pp. 11–22.

[33] Mohammad Sina Karvandi, MohammadHosein Gholamrezaei, Saleh Khalaj Monfared, Soroush Meghdadizanjani, Behrooz Abbassi, Ali Amini, Reza Mortazavi, Saeid Gorgin, Dara Rahmati, and Michael Schwarz. "HyperDbg: Reinventing Hardware-Assisted Debugging". In: *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*. 2022, pp. 1709–1723.

[34] Yuhei Kawakoya, Makoto Iwamura, Eitaro Shioji, and Takeo Hariu. "Api chaser: Anti-analysis resistant malware analyzer". In: *Research in Attacks, Intrusions, and Defenses: 16th International Symposium, RAID 2013, Rodney Bay, St. Lucia, October 23-25, 2013. Proceedings 16*. Springer. 2013, pp. 123–143.

[35] Dhilung Kirat, Giovanni Vigna, and Christopher Kruegel. "Barecloud: Bare-metal analysis-based evasive malware detection". In: *23rd USENIX Security Symposium (USENIX Security 14)*. 2014, pp. 287–301.

[36] Barbara A Kitchenham, David Budgen, and O Pearl Brereton. "The value of mapping studies–A participant-observer case study". In: *14th international conference on evaluation and assessment in software engineering (ease)*. 2010, pp. 1–9.

[37] Barbara Ann Kitchenham, David Budgen, and Pearl Brereton. *Evidence-based software engineering and systematic reviews*. Vol. 4. CRC press, 2015.

[38] Clemens Kolbitsch, Engin Kirda, and Christopher Kruegel. "The power of procrastination: detection and mitigation of execution-stalling malicious code". In: *Proceedings of the 18th ACM conference on Computer and communications security*. 2011, pp. 285–296.

[39] Christopher Kruegel, William Robertson, Fredrik Valeur, and Giovanni Vigna. "Static disassembly of obfuscated binaries". In: *USENIX security Symposium*. Vol. 13. 2004, pp. 18–18.

[40] Alexander Küchler, Alessandro Mantovani, Yufei Han, Leyla Bilge, and Davide Balzarotti. "Does Every Second Count? Time-based Evolution of Malware Behavior in Sandboxes." In: *NDSS*. 2021.

[41] J Richard Landis and Gary G Koch. "An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers". In: *Biometrics* (1977), pp. 363–374.

[42] Kevin Leach, Chad Spensky, Westley Weimer, and Fengwei Zhang. "Towards transparent introspection". In: *2016 IEEE 23rd International Conference on Software Analysis, Evolution, and Reengineering (SANER)*. Vol. 1. IEEE. 2016, pp. 248–259.

[43] Tamas K Lengyel, Steve Maresca, Bryan D Payne, George D Webster, Sebastian Vogl, and Aggelos Kiayias. "Scalability, fidelity and stealth in the DRAKVUF dynamic malware analysis system". In: *Proceedings of the 30th annual computer security applications conference*. 2014, pp. 386–395.

[44] Shijia Li, Chunfu Jia, Pengda Qiu, Qiyuan Chen, Jiang Ming, and Debin Gao. "Chosen-Instruction Attack Against Commercial Code Virtualization Obfuscators". In: *In Proceedings of the 29th Network and Distributed System Security Symposium*. 2022.

[45] Martina Lindorfer, Clemens Kolbitsch, and Paolo Milani Comparetti. "Detecting environment-sensitive malware". In: *Recent Advances in Intrusion Detection: 14th International Symposium, RAID 2011, Menlo Park, CA, USA, September 20-21, 2011. Proceedings 14*. Springer. 2011, pp. 338–357.

[46] Kangjie Lu, Dabi Zou, Weiping Wen, and Debin Gao. "deRop: removing return-oriented programming from malware". In: *Proceedings of the 27th Annual Computer Security Applications Conference*. 2011, pp. 363–372.

[47] Alessandro Mantovani, Simone Aonzo, Yanick Fratantonio, and Davide Balzarotti. "{RE-Mind}: a First Look Inside the Mind of a Reverse Engineer". In: *31st USENIX Security Symposium (USENIX Security 22)*. 2022, pp. 2727–2745.

[48] Jonathan AP Marpaung, Mangal Sain, and Hoon-Jae Lee. "Survey on malware evasion techniques: State of the art and challenges". In: *2012 14th International Conference on Advanced Communication Technology (ICACT)*. IEEE. 2012, pp. 744–749.

[49] Lorenzo Martignoni, Mihai Christodorescu, and Somesh Jha. "Omniunpack: Fast, generic, and safe unpacking of malware". In: *Twenty-Third Annual Computer Security Applications Conference (ACSAC 2007)*. IEEE. 2007, pp. 431–441.

[50] James Mattei, Madeline McLaughlin, Samantha Katcher, and Daniel Votipka. "A Qualitative Evaluation of Reverse Engineering Tool Usability". In: *Proceedings of the 38th Annual Computer Security Applications Conference*. 2022, pp. 619–631.

[51] Grégoire Menguy, Sébastien Bardin, Richard Bonichon, and Cauim de Souza Lima. "Search-based local black-box deobfuscation: understand, improve and mitigate". In: *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. 2021, pp. 2513–2525.

[52] Jiang Ming, Dongpeng Xu, Li Wang, and Dinghao Wu. "Loop: Logic-oriented opaque predicate detection in obfuscated binary code". In: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. 2015, pp. 757–768.

[53] Andreas Moser, Christopher Kruegel, and Engin Kirda. "Exploring multiple execution paths for malware analysis". In: *2007 IEEE Symposium on Security and Privacy (SP'07)*. IEEE. 2007, pp. 231–245.

[54] Anh M Nguyen, Nabil Schear, HeeDong Jung, Apeksha Godiyal, Samuel T King, and Hai D Nguyen. "Mavmm: Lightweight and purpose built vmm for malware analysis". In: *2009 Annual Computer Security Applications Conference*. IEEE. 2009, pp. 441–450.

[55] Anna-Marie Ortloff, Christian Tiefenau, and Matthew Smith. "{SoK}: I Have the (Developer) Power! Sample Size Estimation for Fisher's Exact,{Chi-Squared},{McNemar's}, Wilcoxon {Rank-Sum}, Wilcoxon {Signed-Rank} and t-tests in {Developer-Centered} Usable Security". In: *Nineteenth Symposium on Usable Privacy and Security (SOUPS 2023)*. 2023, pp. 341–359.

[56] *Overview of Qualitative Methods and Analytic Techniques*. https : / / www . nsf . gov / pubs / 1997 / nsf97153/chap_3.htm.

[57] Hernan Palombo, Armin Ziaie Tabari, Daniel Lende, Jay Ligatti, and Xinming Ou. "An ethnographic understanding of software (in) security and a co-creation model to improve secure software development". In: *Proceedings of the Sixteenth Symposium on Usable Privacy and Security*. 2020.

[58] Fei Peng, Zhui Deng, Xiangyu Zhang, Dongyan Xu, Zhiqiang Lin, and Zhendong Su. "X-force: Force-executing binary programs for security applications". In: *USENIX Security symposium 2014*.

[59] Birgit Penzenstadler, Ankita Raturi, Debra Richardson, Coral Calero, Henning Femmer, and Xavier Franch. "Systematic mapping study on software engineering for sustainability (SE4S)". In: *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*. 2014, pp. 1–14.

[60] Kai Petersen, Robert Feldt, Shahid Mujtaba, and Michael Mattsson. "Systematic mapping studies in software engineering". In: *12th International Conference on Evaluation and Assessment in Software Engineering (EASE) 12*. 2008, pp. 1–10.

[61] Kai Petersen, Sairam Vakkalanka, and Ludwik Kuzniarz. "Guidelines for conducting systematic mapping studies in software engineering: An update". In: *Information and software technology* 64 (2015), pp. 1–18.

[62] Meisam Ranjbari, Zahra Shams Esfandabadi, Tetiana Shevchenko, Naciba Chassagnon-Haned, Wanxi Peng, Meisam Tabatabaei, and Mortaza Aghbashlo. "Mapping healthcare waste management research: Past evolution, current challenges, and future perspectives towards a circular economy transition". In: *Journal of hazardous materials* 422 (2022), p. 126724.

[63] Kevin A Roundy and Barton P Miller. "Hybrid analysis and control of malware". In: *Recent Advances in Intrusion Detection: 13th International Symposium, RAID 2010, Ottawa, Ontario, Canada, September 15-17, 2010. Proceedings 13*. Springer. 2010, pp. 317–338.

[64] Paul Royal, Mitch Halpin, David Dagon, Robert Edmonds, and Wenke Lee. "PolyUnpack: Automating the Hidden-Code Extraction of Unpack-Executing Malware". In: *2006 22nd Annual Computer Security Applications Conference (ACSAC'06)*. 2006, pp. 289–300. DOI: 10.1109/ACSAC.2006.38.

[65] Monirul Sharif, Andrea Lanzi, Jonathon Giffin, and Wenke Lee. "Automatic reverse engineering of malware emulators". In: *2009 30th IEEE Symposium on Security and Privacy*. IEEE. 2009, pp. 94–109.

[66] Junfu Shen and Jiang Ming. "Mba-blast: unveiling and simplifying mixed boolean-arithmetic obfuscation". In: (2021).

[67] Yan Shoshitaishvili, Ruoyu Wang, Christopher Salls, Nick Stephens, Mario Polino, Audrey Dutcher, John Grosen, Siji Feng, Christophe Hauser, Christopher Kruegel, and Giovanni Vigna. "SoK: (State of) The Art of War: Offensive Techniques in Binary Analysis". In: *IEEE Symposium on Security and Privacy*. 2016.

[68] Chad Spensky, Hongyi Hu, and Kevin Leach. "LO-PHI: Low-Observable Physical Host Instrumentation for Malware Analysis." In: *NDSS*. 2016.

[69] Deepa Srinivasan, Zhi Wang, Xuxian Jiang, and Dongyan Xu. "Process out-grafting: an efficient" out-of-vm" approach for fine-grained process execution monitoring". In: *Proceedings of the 18th ACM conference on Computer and communications security*. 2011, pp. 363–374.

[70] Prachi Srivastava and Nick Hopwood. "A practical iterative framework for qualitative data analysis". In: *International journal of qualitative methods* 8.1 (2009), pp. 76–84.

[71] Chao Su, Xuhua Ding, and Qingkai Zeng. "Catch you with cache: Out-of-VM introspection to trace malicious executions". In: *2021 51st Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*. IEEE. 2021, pp. 326–337.

[72] Tyler W Thomas, Madiha Tabassum, Bill Chu, and Heather Lipford. "Security during application development: An application security expert perspective". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 2018, pp. 1–12.

[73] Sharath K Udupa, Saumya K Debray, and Matias Madou. "Deobfuscation: Reverse engineering obfuscated code". In: *12th Working Conference on Reverse Engineering (WCRE'05)*. IEEE. 2005, 10–pp.

[74] Xabier Ugarte-Pedrero, Davide Balzarotti, Igor Santos, and Pablo G Bringas. "Rambo: Run-time packer analysis with multiple branch observation". In: *Detection of Intrusions and Malware, and Vulnerability Assessment: 13th International Conference, DIMVA 2016, San Sebastián, Spain, July 7-8, 2016, Proceedings 13*. Springer. 2016, pp. 186–206.

[75] Amit Vasudevan and Ramesh Yerraballi. "Cobra: Fine-grained malware analysis using stealth localized-executions". In: *2006 IEEE Symposium on Security and Privacy (S&P'06)*. IEEE. 2006, 15–pp.

[76] Daniel Votipka, Kelsey R Fulton, James Parker, Matthew Hou, Michelle L Mazurek, and Michael Hicks. "Understanding security mistakes developers make: Qualitative analysis from build it, break it, fix it". In: *29th USENIX Security Symposium (USENIX Security 20)*. 2020, pp. 109–126.

[77] Daniel Votipka, Seth M Rabin, Kristopher Micinski, Jeffrey S Foster, and Michelle M Mazurek. "An observational investigation of reverse engineers' processes". In: *Proceedings of the 29th USENIX Conference on Security Symposium*. 2020, pp. 1875–1892.

[78] Daniel Votipka, Rock Stevens, Elissa Redmiles, Jeremy Hu, and Michelle Mazurek. "Hackers vs. testers: A comparison of software vulnerability discovery processes". In: *2018 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2018, pp. 374–391.

[79] Jeffrey Wilhelm and Tzi-cker Chiueh. "A forced sampled execution approach to kernel rootkit identification". In: *Recent Advances in Intrusion Detection: 10th International Symposium, RAID 2007, Gold Goast, Australia, September 5-7, 2007. Proceedings 10*. Springer. 2007, pp. 219–235.

[80] Carsten Willems, Thorsten Holz, and Felix Freiling. "Toward automated dynamic malware analysis using cwsandbox". In: *IEEE Security & Privacy* 5.2 (2007), pp. 32–39.

[81] Carsten Willems, Ralf Hund, Andreas Fobian, Dennis Felsch, Thorsten Holz, and Amit Vasudevan. "Down to the bare metal: Using processor features for binary analysis". In: *Proceedings of the 28th Annual Computer Security Applications Conference*. 2012, pp. 189–198.

[82] Dongpeng Xu, Jiang Ming, Yu Fu, and Dinghao Wu. "VMHunt: A verifiable approach to partially-virtualized binary code simplification". In: *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. 2018, pp. 442–458.

[83] Zhaoyan Xu, Jialong Zhang, Guofei Gu, and Zhiqiang Lin. "Goldeneye: Efficiently and effectively unveiling malware's targeted environment". In: *Research in Attacks, Intrusions and Defenses: 17th International Symposium, RAID 2014, Gothenburg, Sweden, September 17-19, 2014. Proceedings 17*. Springer. 2014, pp. 22–45.

[84] Babak Yadegari and Saumya Debray. "Symbolic execution of obfuscated code". In: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. 2015, pp. 732–744.

[85] Babak Yadegari, Brian Johannesmeyer, Ben Whitely, and Saumya Debray. "A generic approach to automatic deobfuscation of executable code". In: *2015 IEEE Symposium on Security and Privacy*. IEEE. 2015, pp. 674–691.

[86] Khaled Yakdan, Sergej Dechand, Elmar Gerhards-Padilla, and Matthew Smith. "Helping johnny to analyze malware: A usability-optimized decompiler and malware analysis user study". In: *2016 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2016, pp. 158–177.

[87] Lok-Kwong Yan, Manjukumar Jayachandra, Mu Zhang, and Heng Yin. "V2e: combining hardware virtualization and softwareemulation for transparent and extensible malware analysis". In: *Proceedings of the 8th ACM SIGPLAN/SIGOPS conference on Virtual Execution Environments*. 2012, pp. 227–238.

[88] Miuyin Yong Wong, Matthew Landen, Manos Antonakakis, Douglas M Blough, Elissa M Redmiles, and Mustaque Ahamad. "An inside look into the practice of malware analysis". In: *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. 2021, pp. 3053–3069.

[89] Wei You, Zhuo Zhang, Yonghwi Kwon, Yousra Aafer, Fei Peng, Yu Shi, Carson Harmon, and Xiangyu Zhang. "Pmp: Cost-effective forced execution with probabilistic memory pre-planning". In: *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2020, pp. 1121–1138.

[90] Junyuan Zeng, Yangchun Fu, and Zhiqiang Lin. "Pemu: A pin highly compatible out-of-vm dynamic binary instrumentation framework". In: *Proceedings of the 11th ACM SIGPLAN/SIGOPS International Conference on Virtual Execution Environments*. 2015, pp. 147–160.

[91] Junyuan Zeng, Yangchun Fu, Kenneth A Miller, Zhiqiang Lin, Xiangyu Zhang, and Dongyan Xu. "Obfuscation resilient binary code reuse through trace-oriented programming". In: *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*. 2013, pp. 487–498.

[92] Fengwei Zhang, Kevin Leach, Angelos Stavrou, Haining Wang, and Kun Sun. "Using hardware features for increased debugging transparency". In: *2015 IEEE Symposium on Security and Privacy*. IEEE. 2015, pp. 55–69.

[93] Fengwei Zhang, Kevin Leach, Kun Sun, and Angelos Stavrou. "Spectre: A dependable introspection framework via system management mode". In: *2013 43rd Annual IEEE/IFIP international conference on dependable systems and networks (DSN)*. IEEE. 2013, pp. 1–12.

[94] Qinghua Zhang and Douglas S Reeves. "Metaaware: Identifying metamorphic malware". In: *Twenty-Third Annual Computer Security Applications Conference (ACSAC 2007)*. IEEE. 2007, pp. 411–420.

A Survey Questionnaire

Background and Experience.

- How many years of experience do you have analyzing malware?
- Can you please describe your job role?
- We would like to get to know more about you, please provide your LinkedIn profile. If you do not have a LinkedIn profile, please describe your work experience and education such as your highest level of education and major (if applicable).
- Which of the following best describes your malware analysis objective? Extract easily obtained string based IOCs such as hashes, domain names and IP addresses from malware samples, Focus on identifying potentially malicious activities exhibited by malware samples using network and host artifacts, Perform malware analysis to identify the strategies and intentions behind threat actor's attack campaigns which is accomplished by understanding the Tactics, Techniques and Procedures (check all that apply).

Malware Evasion.

- How do you define an evasive malware sample?
- How often do you analyze evasive malware?
- How do you determine that a malware sample is evasive?
- Which of the following type of analysis do you tend to rely on more when identifying an evasive malware sample? Dynamic Analysis, Static Analysis, I use both static and dynamic analysis equally, Other.
- Which of the following type of analysis do you tend to rely on more when analyzing evasive malware samples? Dynamic Analysis, Static Analysis, I use both static and dynamic analysis equally, Other.
- Do you consider evasive malware to be challenging to analyze? Yes, No, Other.
- What is the biggest challenge when analyzing evasive malware samples?
- Based on your experience, have evasive tactics become more sophisticated over time? If so, how?
- Based on your experience, is there any correlation between the malware family and the evasive tactics?
- Based on your experience, which of the following category of anti-analysis is the most challenging? Anti-disassembly, Anti-debugging, Sandbox Evasion, Other.
- Based on your experience, what are the most common types of sandbox evasion tactics? Delayed Execution: execute malware after a short period of time to successfully leave the sandbox, Environmental Awareness: verify that it is being executed in a real life environment, System Analysis: look for system characteristics like CPU core count and system reboots, User Interaction: detect user actions like mouse clicks and document scrolling, Data Obfuscation: tricks the sandbox by changing the DNS names or encrypting API calls.

- Does your analysis process differ when you are analyzing an evasive malware sample? If so, how?

B Interview Questions

Identifying and Analyzing Evasive Malware.

- At a high level, what is your workflow when analyzing a malware sample?
- When do you begin to consider the possibility that the sample is evasive?
- How do you identify an evasive malware sample?
- What percentage of the malware samples that you analyze are evasive?
- Is there anything that you would like to change in the process of identifying an evasive malware sample?
- In your questionnaire you mentioned that you use both static and dynamic analysis. Can you explain why you use both and what causes you to switch from one to the other?
- Is your workflow documented or standardized?
- How did you come up with your current workflow?
- Do you hold a college degree? If so, what was your major, and did it help with malware analysis?
- Do you know if your workflow is similar to your coworker's workflow in your group?
- Can you walk me through your process of analyzing an already identified evasive malware sample?
- What are you trying to accomplish when analyzing an evasive sample? What information do you want to extract?
- Is understanding evasion tactics helpful or do you just want to bypass them?
- What are the challenges that you encounter in your workflow and what tools would help you overcome them?

Techniques for Handling Evasive Malware.

- How do you handle malware that uses code obfuscation?
- Do you consider fingerprinting a significant evasive tactic? If so, what steps do you take to mitigate it?
- How do you handle malware that employs timing-based evasion techniques?
- How do you handle malware that checks for user interaction? If you don't, why not?
- When doing dynamic analysis, are you concerned with the malware detecting that it's being executed in an analysis environment? If so, what steps do you take to mitigate this?
- Which evasive malware techniques do you consider to be the most challenging to analyze?

- Which evasive malware techniques do you consider to be the most time-consuming to analyze?
- Why do you think evasive malware remains challenging to analyze?

Use of Existing Tools for Malware Analysts.

- What tools do you use when analyzing an evasive malware sample?
- Which tool would you consider to be the most helpful in your analysis workflow and why?
- Which tool do you use but would you like to be improved?
- When was the last time you implemented a new tool into your workflow?
- Where do you find new tools?
- What qualities do you consider when selecting a new tool?
- Is there any limitation or challenge that you would like a tool to automate?

C Codebook

Below is our complete codebook in their corresponding categories. We provide a brief description of the codes that may need further explanation.

Participant Role.

- Job Description: current job title and job
- Experience: years of experience in malware analysis and previous jobs.
- Education

Organization.

- Escalation: When malware samples get sent between teams.
- Mentoring: If malware analysts mention having a mentor or mentoring another individual regarding malware analysis.
- Operational Analyst: Participants that mention they work at AV companies, as malware analysts, or as incident responders.
- Research Focus: When the analysts' objective reaches beyond detection and requires a more in-depth understanding of the malware's origin and purpose.
- Restrictions: Factors that restrict the analysis process.
- Resources Influence Workflow: When malware analysts' workflow is affected or benefited by specific resources within their organization.

Malware.

- Acquisition of Malware: the process of acquiring a potentially malicious program.
- Malware Context: information the analysts receive about the malware sample.

- Commodity Malware: any information mentioned about commonly viewed malware.
- Sophisticated Malware: any information mentioned about sophisticated malware.
- Example of Malware: analyst mentions specific examples of malware samples, including the name of the family.
- Targeted Malware: when the analyst mentions a malware sample that avoids detection unless it is executed in its desired environment.
- Multiple Stages: analysts describes multi-stage malware
- Type of Malware Informs Analysis Process: a case when a participant's process changes based on the malware sample.
- Programming Language

Analysis Workflow.

- High-Level Workflow: description of the analysis workflow used by analysts to analyze any malware sample.
- Switch Trigger Between Static and Dynamic Analysis: what causes analysts to go back and forth between static and dynamic analysis.
- Need for Automation: a process that analysts mention would benefit from automation.
- Standardized Workflow: if the analyst mentions they have a standardized analysis workflow.
- Non-Standardized workflow: if the analyst mentions they do not have a standardized analysis workflow.
- Suspicious Activity: signs that may provide analysts hints about the program's malicious behavior.
- Process of Generating Signatures: how malware analysts create signatures.
- Objectives: the objectives analysts have when analyzing a malware sample.
- Automated Triage: a malware analysis pipeline through which each sample undergoes automated processing.
- Useful Information from One Type of Analysis to the Other: specific information that analysts take from static analysis to use during dynamic analysis or vice versa.

Static Analysis.

- Static Analysis Preference
- Basic Static Analysis: static analysis process such as checking for strings.
- Advanced Static Analysis: how analysts analyze the malware binary in a disassembler.
- Benefits of Static Analysis
- Limitations of Static Analysis
- Locating Suspicious Activity: how malware analysts locate suspicious activity in a disassembler.
- Where They Begin: how do analysts begin analyzing a sample in a disassembler.

Dynamic Analysis.

- Dynamic Analysis Preference
- Basic Dynamic Analysis: sandbox execution
- Advanced Dynamic Analysis: debugging process
- Sandbox configuration
- Sandbox Limitations
- Benefits of Sandbox
- Multiple Sandbox Execution: if they execute the sample multiple times and why
- Symbolic Execution: the use of symbolic execution
- Contributing to Sandbox: when analysts use new information from their analysis to improve sandboxes.
- Unpack Sample: process of unpacking a malware sample
- Use of Sandbox Report
- Bare metal: execute malware in bare metal systems.
- Emulation: analysts use emulation techniques.

Malware Evasion.

- Definition of Evasion
- Detecting Evasive Malware
- Frequency of Evasive Malware
- Most Challenging Evasive Techniques
- Most Time-Consuming Evasive Techniques
- Most Common Evasive Techniques
- Why Analyzing Evasive Malware Remains Challenging
- Correlation Between Malware Family & Evasive Technique
- Purpose of evasion technique
- Layers of Evasion
- Evasion-Based Signatures

Static Analysis Evasion.

- Static Analysis Evasion Techniques
- Frequency of static analysis evasion
- Bypassing Static Analysis Evasion: how malware analysts overcome static analysis evasion techniques.

Dynamic Analysis Evasion.

- Dynamic Analysis Evasion Techniques
- Frequency of Dynamic Analysis Evasion
- Bypassing dynamic analysis evasion: how malware analysts overcome dynamic analysis evasion techniques.

Implementation.

- Discovering New Tools: where analysts find new tools.
- Does Not Use New Tools Often: when analysts mention that they do not use new tools often.
- Willingness to Implement New Tools

- Qualities: when analysts mention qualities that they consider when deciding to use a new tool.

Tools.

- Malware Analysis Tools
- Most Helpful Tool
- Improvements for Tools: when analysts mention how existing tools can be improved.
- New Tool Idea: when analysts mention ideas for new tools.
- Internal Tools: when analysts mention tools made exclusively for their organization.
- Custom Scripts: when malware analysts describe how they create custom scripts.
- Open Source Vs. Custom Tools: when analysts compare open source with tools internal to the organization.
- Hard to Apply in Practice: when analysts mention that certain approaches are difficult to apply in practice.

D   Participants

ID	Education	Yrs.	Analysis Preference
P1	N/A	10	Static Analysis
P2	IT	7	Static Analysis
P3	N/A	8	Static Analysis
P4	Math	10	Static Analysis
P5	CS	15	NA
P6	CE	11	Static Analysis
P7	IT	15	NA
P8	N/A	3	Static Analysis
P9	Computer Networks	3	NA
P10	Information Assurance	11	Dynamic Analysis
P11	CS & Math	8	Static Analysis
P12	CS	9	Static Analysis
P13	CS	6	Static Analysis
P14	CS	7	NA
P15	EE & Math	9	Static Analysis
P16	CS	27	NA
P17	CS	9	Static Analysis
P18	IT	5	NA s
P19	CS	12	Static Analysis
P20	CS	12	NA
P21	Digital Forensics	8	Static Analysis
P22	EE & Math	15	Dynamic Analysis
P23	CE	12	Static Analysis
P24	N/A	15	Dynamic Analysis

Table 3: Participants' Education (IT: Information Technology, CS: Computer Science, EE: Electrical Engineering, CE: Computer Engineering), Years of Experience, Analysis Preference, and Tiers [88] (Behavior: Goal of identifying potentially malicious activity, TTPs: Goal of understanding tactics, techniques, and procedures).

Write, Read, or Fix?
Exploring Alternative Methods for Secure Development Studies

Kelsey R. Fulton[*♡], Joseph Lewis[◇], Nathan Malkin[†♡], Michelle L. Mazurek[◇]
[*]*Colorado School of Mines;* [◇]*University of Maryland;* [†]*New Jersey Institute of Technology*

Abstract

When studying how software developers perform security tasks, researchers often ask participants to write code. These studies can be challenging because programming can be time-consuming and frustrating. This paper explores whether alternatives to code-writing can yield scientifically valid results while reducing participant stress. We conducted a remote study in which Python programmers completed two encryption tasks using an assigned library by either writing code from scratch, reading existing code and identifying issues, or fixing issues in existing code. We found that the read and fix conditions were less effective than the write condition in revealing security problems with APIs and their documentation, but still provided useful insights. Meanwhile, the read and especially fix conditions generally resulted in more positive participant experiences. Based on these findings, we make preliminary recommendations for how and when researchers might best use all three study design methods; we also recommend future work to further explore the uses and trade-offs of these approaches.

1 Introduction

Secure software development is a difficult task, as demonstrated by the many vulnerabilities discovered in production code on a regular basis [12, 29, 35]. Key causes of these vulnerabilities include developers failing to use the right tools or resources [4, 19, 31, 32, 42], making mistakes when writing

---

[♡]Work done while at University of Maryland

code [52], or fundamentally misunderstanding necessary and important security concepts [19, 31, 52]. Addressing these challenges requires better understanding them; to wit, studying how developers approach secure development and how and why errors occur.

How best to conduct these studies, however, remains in some respects an open challenge. Many secure-development studies rely on code-writing tasks in order to observe developers' processes, decisions, and missteps [3,19,30,41,52]. These studies produce valuable results, but they can be very challenging to conduct: code-writing tasks are time-consuming and difficult to scope narrowly for lab experiments, and ecological validity is challenging because professional software development environments are hard to simulate [4, 40]. Further, the specialized population of software developers can be challenging to recruit and retain: they are difficult to reach, and they participate in studies outside their normal work hours, often for hourly rates much lower than their regular pay for software engineering [31, 45]. Frustration while participating in such studies can lead to high dropout rates, resulting in smaller, less powerful sample sizes [2].

Given this context, it is imperative for secure-development researchers to understand whether there are alternative approaches to conducting code writing studies that—at least for some research questions and contexts—will yield similar, scientifically valid results while reducing the stress and frustration for participants and researchers. As a first step to address this problem, Danilova et al. explored the possibility of substituting code review for code writing: participants wrote code reviews about snippets from a prior study about secure password storage [16]. The results indicated that code reviewing studies are potentially useful in addressing certain types of secure-development research questions.

Here, we build on this result with another exploratory study, this time investigating both reading and fixing insecure code as methodological alternatives to code writing. In contrast to Danilova et al., we directly compare these methods, measuring both secure-development outcomes and participants' experiences taking part in the study.

Specifically, we conducted a remote experimental study in which 112 Python programmers completed two symmetric encryption tasks (generating/storing a key and encrypting/decrypting data) using one of two encryption libraries. These study tasks were modeled on an earlier study by Acar et al. that measured the usability of various Python encryption APIs [2]. Participants were each assigned to one of three methodological conditions: writing secure code from scratch (*Write*), reading existing code and finding and explaining vulnerabilities and bugs (*Read*), or finding and fixing vulnerabilities and bugs in existing code (*Fix*). Participants then completed a survey about their experience taking part in the study.

Using this study, we address two key research questions:

RQ1 Do the *Read* and *Fix* conditions provide the same results about functionality and security as the *Write* condition?

RQ2 Do participants in the *Read* and *Fix* conditions experience fewer negative effects (drop-out rate, frustration, time to complete) than those in the *Write* condition?

We found that *Read* and *Fix* were less effective than *Write* in revealing security problems with APIs and their documentation. Participants in the *Fix* condition were particularly likely to focus only on getting code to run and pass provided tests; as such, they identified a much narrower set of vulnerabilities than *Read* participants. However, when they did identify vulnerabilities, *Fix* participants' attempts to remediate them did reveal interesting shortcomings in API documentation. Both the *Read* and *Fix* conditions provided insight into the kinds of vulnerabilities participants could recognize and remediate, in lower- and higher-effort scenarios respectively.

Overall, participants in the *Read* and *Fix* conditions reported more positive study experiences than *Write* participants, including more fun and less frustration. While participants in the *Read* condition spent slightly less time working on the study than those in the *Write* condition, participants in the *Fix* condition spent substantially less time.

Based on our qualitative and quantitative observations, we make preliminary recommendations for how and when researchers might best use all three study-design methods; we also recommend future work to further explore the uses and trade-offs of these approaches. Our exploratory study tests one possible implementation of *Write*, *Read*, and *Fix* study designs, in one experimental context aimed at answering a specific question about comparing APIs. As such, it cannot fully answer all questions about how and when it may be appropriate to deploy these methods. Nonetheless, we believe our study makes a valuable contribution to the ongoing evolution of best practices for empirical studies of secure development.

2   Related work

Prior research has explored how to recruit developers for studies and measure their efforts.

Participant recruitment. Prior work has studied the validity of varying recruitment approaches. Yamashita et al. explored the use of freelance marketplaces to recruit participants, concluding that while these services offer flexibility, low cost, and access to a wide population, there is uncertainty about developers' background and skills [53]. Baltes et al. evaluated the use of various sampling methodologies for software development studies, finding that sampling through public media yielded the best results [9]. Acar et al. explored the use of GitHub as a recruitment tool for studies by recruiting 307 active GitHub users to complete security-related programming tasks, finding no statistical difference for functionality or security between participants that were students and professionals [5]. To provide insight into the ecological validity of using computer science students in developer studies, Naiakshina et al. recruited professional software developers to complete programming problems from a prior study [31] and compared the results [30], finding that developers performed better than students and freelancers, but the treatment effects, such as prompting for security, held the same for developers and students. Similarly, Salman et al. discovered that students and professionals do not produce substantially different results in software studies [43].

Most recently, researchers have compared and contrasted a variety of recruitment venues, finding that crowdsourcing platforms require screening to get high-quality participants and that CS students at a variety of educational stages are a viable alternative for developers [25, 48]. To understand why professional developers may or may not participate in security studies, Serafini et al. conducted interviews with 30 developers and found that the length of the study, the topic of the study, compensation, and trust in the researchers had an impact on participation [45]. Interviewees were also concerned about their performance on security tasks.

Measuring developers' skill. Prior work has explored the construction of scales and survey questions to evaluate the skills of participants in software developer studies. Feigenspan et al. constructed a scale to measure programming experience of study participants by using questions from published studies that evaluated programming experience [17]. Comparing the participants' answers, they found self-efficacy to be an effective way to measure programming experience. Similarly, Bergersen et al. built a scale to measure programming experience by having 65 professional developers complete 19 Java programming exercises [10].

To aid in recruitment, Danilova et al. expanded this idea by building a screening questionnaire to help filter participants in programming studies, concluding with 6 recommended questions [15]. In follow-on work, they explored the use of time limits to increase the efficiency of screening questionnaires, concluding that implementing a time limit saves time and money while maintaining validity [13]. Focusing on measur-

ing security experience, Votipka et al. built a 15-item scale to measure the security self-efficacy of software developers [51].

Study design for secure development studies. Recent work has explored the use of a priori power calculations in the design of secure development studies, finding that many prior developer studies were underpowered to detect large effects [36].

Other work has explored the design of tasks provided in these studies. To explore the use of task deception, Naiakshina et al. had 40 students complete a password storage task, finding that priming participants for security had a statistically significant effect [33]. In a replication, Danilova et al. found that deception may not be necessary for ecological validity [14]. As an alternate approach to traditional in-person lab studies, Huaman et al. built a virtual study environment to allow researchers to conduct lab studies remotely, while maintaining ecological validity [24]. Most similar to our work, Danilova et al. evaluated the use of code review as a method for secure development studies, finding that code review could be a viable method for future developer studies, but recommending that more research be done into this alternative [16].

3 Method

To understand how *Read* and *Fix* compare to *Write*, we conducted a remote experimental study with Python programmers in which they completed two secure development tasks.

3.1 Study design

To ensure the validity of our study, we decided to partially replicate a prior study. This allowed us to compare the results of our *Write* with the original study's results, while also allowing us to compare the results of our *Write*, *Read*, and *Fix* conditions to each other. We chose to partially replicate a 2016 study from Acar et al. exploring the usability of various Python cryptography APIs [2] because it offered self-contained, short code writing tasks with results that lent themselves well to being compared across different conditions.

The original study [2] used five libraries and two types of encryption (symmetric and asymmetric encryption), for a total of 10 conditions. Comparing three different methods would triple this to 30 conditions, requiring a sample size that would not be feasible when recruiting and compensating developers. (The original study did not compensate participants.) As such, we opted to only replicate a subset of the original conditions (two libraries, one type of encryption), crossed with our three methods, for a total of six conditions.

We used the symmetric encryption tasks from the original study, as they served as the baseline for all of the regressions performed in the original paper, and participants in the symmetric condition produced more functional solutions. Participants were assigned to work in either PyCrypto or Cryptography.io. We selected PyCrypto as it was the baseline library

for all the models in the original paper. We selected Cryptography.io as it proved to be significantly better in regards to security in the original paper and was designed with usability in mind. To facilitate replication, we used the versions of each library used in the original study, because many insecure defaults and security issues identified in the original study have been addressed in modern versions of the libraries. Finally, participants were assigned to one of three experimental conditions: writing code from scratch (*Write*); reading existing code, determining its correctness, and describing anything they wanted to change (*Read*); or reading existing code, determining its correctness, and fixing any vulnerabilities or bugs in the code (*Fix*).

Task selection. Participants were randomly assigned to one of the six conditions and tasked with completing an encrypt/decrypt task and a key generation and storage task. The order they were presented in was randomized.

For the *Write* condition, participants were given stub code and asked to write code that completed the described task. For the encrypt/decrypt task, this meant writing code that encrypted or decrypted plaintext or ciphertext, respectively, using a provided key. For the key generation and storage task, participants were tasked with creating an encryption key (file key) using a provided password, using the file key to encrypt and store another encryption key (task key) to a file in a provided directory name, and recovering the task key from the same file. They were provided with tests for each function and a set of cumulative tests at the very end. The tests covered the basic functionality of each task and ensured that the code ran without failure and returned the correct value. The key generation and storage function stub, encrypt/decrypt function stub, and provided tests for both tasks can be found in the Supplementary Materials.[1]

For the *Read* condition, participants were given already-completed code and asked to read it, determine its correctness, and add comments to identify what they would change and how. For the *Fix* condition, participants were given already-completed code and asked to read it, determine its correctness, and fix it if necessary. In both the *Read* and *Fix* conditions, the code provided contained four unique functionality bugs and one unique security vulnerability for the encrypt/decrypt task and two unique functionality bugs and four unique security vulnerabilities for the key generation and storage task. The functionality bugs ranged in type from those that would cause the code to crash when run to those that would not cause the provided tests to fail. The security vulnerabilities were based on vulnerabilities identified in the original paper [2], to best allow for comparing results between studies. A full list of the functionality bugs and security vulnerabilities in the provided code can be found in Table 1. Participants in the

---

[1]Supplementary materials, as well as a version of the paper with appendices included, can be found at https://osf.io/2nb3g/?view_only=8f83b46a6084440783d88c12e225a46c.

Task	Type	Vuln/bug	Description	Tests fail
Encrypt/decrypt	Security	Fixed IV	IV was set to static value	No
	Functionality	Return plaintext data	Return plaintext instead of ciphertext	Yes
	Functionality	Encrypt key instead of plaintext	Send key to encryption function instead of plaintext	Yes
	Functionality	Use plaintext as encryption key	Use plaintext as key to create cipher object	No
	Functionality	Return ciphertext data	Return ciphertext instead of plaintext	Yes
Key generation and storage	Security	Fixed salt	Salt was set to static value	No
	Security	Weak KDF	PBKDF1 used	No
	Security	Weak hash algorithm	Sha1 used	No
	Security	Bad mode selection	Insecure mode used to encrypt key	No
	Functionality	Wrong name to open	Send wrong variable to open command when writing key to file	Yes
	Functionality	Incorrect length for password check	Use wrong value to check length of password before using in keygen	No

Table 1: Vulnerabilities and bugs that were included in read and fix code snippets

Read and *Fix* conditions were provided with the same tests as participants in the write condition for reference (*Read*) or to test their code (*Fix*). All code and tests can be can be found in the Supplementary Materials.

Study environment. Our participants completed the study remotely through *NERDS* [26], a study environment based on the Developer Observatory [47]. *NERDS* is a customized Jupyter notebook environment that allows participants to complete coding tasks remotely, allowing for easier recruitment.

Once participants consented, they were taken to the instructions for the study. From there, they were able to begin participating. Participants in the *Write* and *Fix* conditions were able to run their code as often as they liked. Participants in the *Read* condition were not provided with a run button for their code, but they were, initially, still able to use hot keys to run code in Jupyter notebooks. These participants' solutions were removed from the study, and the interface was updated to remove this "feature." Participants were able to move forwards or backwards through the tasks at any point. Once participants completed all the tasks, they clicked a *finish* button that took them to the final survey. An example of our study infrastructure can be found in the Supplementary Materials.

Exit survey. Once participants indicated that they were finished with the tasks, they were directed to the final survey. We first asked participants about the perceived security and correctness of their solutions, how frustrating, fun, tedious, and challenging they found the tasks, and what was easy and hard about the tasks. The next section contained the Secure Software Development Self-efficacy Scale, which measures a person's perceived ability to complete various security tasks [51]. The survey concluded by asking participants about their security experience, general development experience, Python development experience, and demographics. Appendix A.1 contains the full survey.

3.2 Data analysis

We analyzed the data we collected using a mix of qualitative and quantitative analysis, which we describe below.

Qualitative analysis. To determine what vulnerabilities and bugs were introduced by participants in *Write* and not identified or fixed by participants in *Read* or *Fix*, we manually reviewed the submissions following processes from prior work [19, 52]. For the *Write* condition, two authors reviewed both tasks to identify any bugs or vulnerabilities present, using the vulnerabilities identified in the original Acar et al. study as a reference point. Each bug or vulnerability was labeled for type (functionality or security) and the specific vulnerability. In addition, we categorized the vulnerabilities into "issue" classes based on the classifications used in prior work [52].

For *Read* and *Fix*, the two authors reviewed submissions using a list of the bugs and vulnerabilities they knew to be present. For *Fix*, vulnerabilities or bugs included in the study setup were labeled for *both* whether participants were able to correctly identify and fix it. For the *Read* condition, vulnerabilities or bugs included in the study setup were labeled for whether participants were able to correctly identify the existence of the vulnerability/bug and whether the changes they proposed would fix the vulnerability/bug, how the participant said they would change the code to address the vulnerability/bug, and why. For both the *Read* and *Fix* conditions, we also labeled any additional issues identified by participants that were not actual bugs or vulnerabilities (false positives) and whether the unneeded fix introduced any new problems.

IRR was calculated for all variables using Krippendorff's α statistic, a conservative measure that considers coders' agreement as an improvement over randomly guessing. We met the recommended threshold for Krippendorff's α of 0.8 [22]. Prior to agreement, all vulnerabilities and bugs were confirmed by both coders, and consensus for all codes was reached through discussion. The final codebook and associated IRR values are in the Supplementary Materials.

Statistical comparisons. To compare our results among conditions and to the results from Acar et al., we performed logistic regressions to explore the impact of the library used on security and functionality (binary outcomes). We added a random intercept to account for multiple tasks from the same participant and used PyCrypto as the baseline, mirroring the regressions in Acar et al. [2]. To understand the impact of the condition on participants, we applied a linear regression for numeric outcomes (time spent), a poisson regression for count outcomes (number of vulnerabilities and bugs), and an ordinal logistic regression for Likert-scale outcomes (reported frustration/fun). For all regressions, the baseline was the *Write* condition to allow for better comparison of our new conditions to the more established experimental method.

3.3 Recruitment

We recruited from Upwork [1], an online freelancing platform, and computer science student mailing lists at multiple universities from May 2022 to July 2023, following best practices [25, 48]. We opted for multiple recruitment approaches to maximize the number of participants from a traditionally challenging population. Given that students and freelancers offer comparable conclusions [25, 30, 31, 48] and the original Github recruitment approach is no longer available, this mixed population proved effective.

Upwork allows researchers to filter participants based on their skillset. We filtered participants for experience with Python and age (18 or older). For the full study, participants were invited if, from their profile, they had completed at least one small project in Python. For recruitment from CS student mailing lists, we created a short screening survey to ensure that participants in the main study would have programming and Python experience. We started with a few questions to understand their general programming experience and occupation and concluded by using the questionnaire created by Danilova et al. [15] to determine if a participant actually had programming experience. A full copy of the screening survey can be found in Appendix A.2. We exclusively invited participants who had Python experience and were able to correctly answer all of the questions from the Danilova et al. questionnaire. Only 23/188 pre-screened participants failed the Danilova measure.

Participants were not compensated for the screening survey, but all main study participants were compensated $35 for completing the study with a possibility for a $5 bonus if participants were able to identify a majority (75%) of the functionality bugs present in the code. This bonus was meant to encourage people to give their best effort when participating in the study and focus on finding all issues within the code, rather than just the obvious issues. We framed the bonus as a reward for meeting a high correctness threshold (without specifying a number of bugs and only counting functionality bugs). Many participants (68% total, 58% of *Fix*, 81% of *Read*,

and 94% of *Write*) received this bonus. While this may have impacted participant interaction, we deemed this important to promoting ecological validity. We discuss this further in Section 5. We discarded any responses where participants skipped all tasks, but kept responses that did not receive the bonus but where some attempt was made, as this gives us valuable insights into participant behavior.

3.4 Ethics and consent

Both surveys (pre-screen and final) and the full study were approved by University of Maryland's and Colorado School of Mines' Institutional Review Board. We obtained informed consent before the pre-screen survey and again before the full study. Participants were informed that they could skip any task or question and drop out of the study at any time.

3.5 Limitations

A key limitation of our work is the age of the study that we are replicating. The cryptographic libraries and their documentation have changed drastically since the original study. This means that any search of online materials for assistance would likely result in information that does not match the version of the library used in our study. We provided participants with a version of the documentation that matched the versions of libraries they were using, and we encouraged them to use the documentation as much as possible. Since we are not actually concerned with evaluating the current usability or security of the libraries, but rather with understanding the effect of the method used to study them, this limitation does not reduce the validity of our results.

The original study aimed to understand the usability of cryptography APIs for professional developers. We aim to semi-replicate this study. However, about half of our participants for this study were students, differing from the original study which used GitHub developers (a practice that is no longer allowed). While students often have less experience than professionals, several recent studies conclude they can be adequate substitutes in secure software development studies [25, 30, 31, 48], as many skilled professionals have limited experience with security specifically [3, 6, 11, 21, 23, 28, 42, 46, 52].

Our goal in this study was to understand the feasibility of using reading and/or fixing code as experimental substitutes for writing code in secure software development studies. This study serves as a single data point in this exploration. We explore whether reading and/or fixing code works to compare the usability of cryptography APIs. Additionally, the tasks in this study were deliberately small and self-contained to allow for easy comparison among the experimental conditions. Thus, our results may not generalize to all kinds of secure software development studies, such as those exploring other secure development issues or tasking participants with build-

	Upwork	CS mailing list
Total participants	*n*=76	*n*=36
Programming experience	7.4 years	5.7 years
Python experience	4.5 years	3 years
Professional programming experience	4.2 years	2.4 years
Professional Python experience	2.5 years	0.8 years
Security experience	1.4 years	0.8 years
Above-average security knowledge	68%	72%
SSD-SES total	46.3	40.8
SSD-SES Vulnerability	26.5	23.6
SSD-SES Communication	19.7	17.2

Table 2: Participant demographics

ing larger projects. However, we believe this study is a good first step toward exploring this phenomenon.

A final possible limitation of this work is the reliance on some self-report data to measure negative experiences of participants, such as reported frustration or fun. While self-report data can be biased or inaccurate, this is the best proxy we have for measuring frustration levels. We attempt to mitigate some of this self-report bias by also collecting other measures of negative effects, such as time spent or dropout rates, and consider these together as a measure for the negative experiences of our participants.

4 Results

In this section, we discuss our participants, our *Write* condition results as compared to the results from Acar et al. [2], and our *Read* and *Fix* compared to our *Write* condition results.

4.1 Participants

In total, 141 participants started our study, with 41 in the *Write* condition, 54 in the *Read* condition, and 46 assigned to the *Fix* condition. In total, 127 participants completed our study with 35 in *Write*, 48 in *Read*, and 44 in *Fix*. However, we removed 15 participants for a variety of reasons such as running or editing code in the *Read* condition prior to removing the use of hot keys (N = 9), skipping every task in the study (N = 5), and not understanding what to do (N = 1). This left us with 112 participants who completed the study; 35 in the *Write* condition, 37 in the *Read* condition, and 40 in the *Fix* condition. Details of participants assigned to each experimental condition and library can be found in Table 3.

Demographics. In general, our participants trended heavily toward male (80%), young (with ages ranging from 18 to 43 and 91% of participants being younger than 40), and educated (65% had at least a bachelor's degree). Our participants came from a variety of ethnic backgrounds, with a plurality identifying as Asian (48%).

On average, our participants had 6.8 years of programming experience and 4 years of Python experience. About 61% of our participants were employed in a professional role that required programming, with the most common job roles being developer and engineer. Of that 61%, 59% used Python in their job. Our participants had 3.8 years of professional programming experience and 2.1 years of professional Python experience. Finally, our participants had fairly little security experience, with an average of 1.2 years of security experience. However, our participants were self-confident in their security abilities, with 72% rating their security knowledge as at least average. While the experience of our participants may seem unusually high for a population including students, among education levels reported by final CS mailing list participants, 20 were consistent with being in college and 13 with being alumni or grad students (N = 36). At the institutions we recruited from, many undergrads enroll with significant high-school programming experience, so the high years of experience for our participants are not all that surprising. Demographics for each recruitment venue can be found in Table 2.

4.2 Replicating results from Acar et al.

First, we compare the *Write* condition to the functionality and security results from Acar et al. [2] (see Table 3).

Functionality. We considered participants' solutions to be functional if they ran, passed the provided tests, and completed the assigned task. If a participant skipped a task, the result was considered not functional. In Acar et al., participants were able to generate slightly more functional solutions using Cryptography.io than with PyCrypto, though this result was not statistically significant. In our study, slightly more PyCrypto participants than Cryptography.io participants produced functional solutions; this difference was likewise not statistically significant (Table 4).

Security.
For solutions deemed functional, we examined their security. In Acar et al., participants were able to generate significantly more secure solutions using Cryptography.io than PyCrypto. We similarly see significantly more secure solutions from participants using Cryptography.io than PyCrypto in our study (Table 3). Participants using Cryptography.io were $4.7\times$ more likely to generate a secure solution (Table 4).

Comparing the security between the encrypt/decrypt and key generation and storage tasks, participants in the original study were most likely to produce a secure solution for the encrypt/decrypt task. We see a similar trend in our study (Table 3), although not as pronounced, with participants producing more secure solutions for the encrypt/decrypt task than the key generation and storage task (17 vs 10 solutions).

The distribution and types of vulnerabilities we found also followed the original study closely (see Table 5). In the en-

	Write			**Read**			**Fix**			**Acar et al. [2]**		
	P[1]	C[2]	T[3]	P	C	T	P	C	T	P	C	T
Started	21	20	*41*	26	28	*54*	22	24	*46*	136	136	*272*
Completed	18	17	*35*	24	24	*48*	21	23	*44*	48	48	*96*
Valid	18	17	*35*	19	18	*37*	21	19	*40*	41	39	*80*
Functionality	83%	59%	–	47%	31%	–	60%	84%	–	85%	90%	–
Key gen/storage	16	9	*25*	6	6	*12*	16	17	*33*	80%	80%	–
Encrypt/decrypt	14	11	*25*	12	5	*17*	9	15	*24*	90%	98%	–
Security	43%	70%	–	50%	36%	–	32%	31%	–	15%	70%	–
Key gen/storage	6	4	*10*	3	2	*5*	3	4	*7*	5%	30%	–
Encrypt/decrypt	7	10	*17*	6	2	*8*	5	6	*11*	20%	100%	–
Time (mins)	–	–	*38.2*	–	–	*30.2*	–	–	*22.5*	–	–	–

[1] PyCrypto   [2] Crypto.io   [3] Total

Table 3: Number of participants, across various conditions of interest. Percentages represent the share of functional/secure solutions among all/functional solutions. We report these percentages for consistency with the original paper.

		Write			**Read**			**Fix**		
Regression	Factor	O.R.	C.I.	*p*	O.R.	C.I.	*p*	O.R.	C.I.	*p*
Functionality	Cryptography.io	0.1	[0.0, 1.2]	0.082	0.5	[0.1, 1.7]	0.261	4.2	[1.3, 24.8]	0.034*
Security	Cryptography.io	4.7	[1.4, 19.4]	0.017*	3.0	[0.2, 55.9]	0.954	0.4	[0.0, 4.3]	0.430

Table 4: Final logistic regression for effect of library on functionality and security in each condition.

crypt/decrypt task, they matched exactly: the most common vulnerability was using a static initialization vector, followed closely by using a weak encryption mode and using a weak encryption algorithm. There were only slight differences between the studies in the key generation and storage tasks: in ours, the most common vulnerability was storing the key unencrypted, followed closely by failing to use a key derivation function (KDF), using a custom key derivation function, and using a static salt. In the original, the top three included using an insecure encryption mode (instead of the custom KDF).

4.3 Comparing functionality among conditions

Next, we compare the overall functionality results and the specific functionality bugs introduced, identified, and fixed, among our three conditions. Throughout this section, we use B_w to represent the number of bugs from the *Write* condition and, analogously, B_r for *Read*, and B_f for *Fix*.

Similar to *Write*, we considered a solution in the *Fix* condition to be functional if it ran, passed all the tests, and completed the task. For the the *Read* condition, we considered the code to be functional if the participant identified and correctly addressed all the functionality bugs we introduced, as described in Table 1. Specific counts for functional solutions per condition and task can be found in Table 3.

Overall, participants in *Write* and *Fix* were able to produce more functional solutions than those in *Read* for both libraries. This is likely because participants in *Read* were unable to run and test their code, which made identifying bugs difficult.

Comparing the two libraries to each other, in the *Read*

condition, participants using PyCrypto were able to produce more functional solutions than those using Cryptography.io (47% vs 31%). This mirrors the result for the *Write* condition, discussed in Section 4.2 above. Conversely, participants in the *Fix* condition were more likely to produce a functional solution using Cryptography.io than PyCrypto (84% to 60%). However, none of these differences in any condition were statistically significant (Table 4).

Looking at the individual tasks, *Read* participants produced more functional solutions for the encrypt/decrypt task than the key generation and storage task. This again mirrors the result for *Write* as well as the original study. Conversely, in the *Fix* condition, participants produced more functional solutions for the key generation and storage task. We hypothesize that this relates to the specific functionality issues we inserted in *Fix*: participants in this condition appeared to prioritize passing the provided tests, which did not flag the functionality bug we inserted into the encrypt/decrypt task.

Examining bugs in *Write*, *Read*, and *Fix*. In total, participants introduced 34 bugs in the *Write* condition, left 81/222 bugs unidentified in the *Read* condition, and left 70/240 bugs unidentified in the *Fix* condition. In the *Write* and *Read* condition, participants introduced or left more bugs when using Cryptography.io ($B_w = 25$, $B_r = 49$) than when using PyCrypto ($B_w = 9$, $B_r = 32$). In the *Fix* condition, participants using PyCrypto ($B_f = 38$) introduced substantially more bugs than those who used Cryptography.io ($B_f = 19$).

Participants focus on 'test-centric' bugs in *Write* and *Fix*.

Function	Issue	Vuln	Write			Read			Fix			Acar et al. [2]		
			P[1]	C[2]	T = 59[3]	P	C	T = 110	P	C	T = 176	P	C	T
Encrypt/decrypt	*Static value*	Static IV	5	0	5	13	6	19	14	16	30	29	0	29
		Total	5	0	5	13	6	19	14	16	30	29	0	29
	Weak choice	Insecure alg	3	1	4	–	–	–	–	–	–	17	0	17
		Insecure mode	4	0	4	–	–	–	–	–	–	23	0	23
		Total	7	1	8	–	–	–	–	–	–	40	0	40
	Total		12	1	13	13	6	19	14	16	30	69	0	69
Key gen/storage	*No encryption*	Key plain	3	7	10	–	–	–	–	–	–	4	7	11
		Total	3	7	10	–	–	–	–	–	–	4	7	11
	Static value	Static IV	2	0	2	–	–	–	–	–	–	3	0	3
		Static salt	2	3	5	11	9	20	19	17	36	1	10	11
		Static key	2	2	4	–	–	–	–	–	–	–	–	–
		Total	6	5	11	11	9	20	19	17	36	4	10	14
	Weak choice	No KDF	6	3	9	–	–	–	–	–	–	15	1	16
		Custom KDF	5	0	5	–	–	–	–	–	–	0	0	0
		Weak KDF	–	–	–	11	–	11	20	–	20	1	0	1
		Weak hash	–	–	–	–	9	9	–	17	17	0	0	0
		KDF iter	3	0	3	14	11	25	16	20	36	2	0	2
		Insecure alg	3	0	3	–	–	–	–	–	–	11	0	11
		Insecure mode	5	0	5	15	11	26	19	18	37	14	0	14
		Total	22	3	25	40	31	71	55	55	110	43	1	44
	Total		31	21	46	53	37	91	74	72	146	51	18	69

[1] PyCrypto [2] Crypto.io [3] Total

Table 5: Number of vulnerabilities for each issue and the number of projects each vulnerability was introduced in.

For most functionality bugs in the *Write* condition, participants' code did run and pass the provided tests; however, it did not complete the required task. For example, the most common functionality issue in *Write* was caused by failing to store the encryption information correctly ($B_w = 6$), such as failing to store the key in the provided directory, as per the instructions ($B_w = 4$), or not storing the key in a file ($B_w = 2$).

Similarly, most of the bugs unidentified in the *Fix* condition were caused by failing to complete the task rather than the code failing to run or pass the tests ('test-centric' bugs). The most common functionality issue left unidentified in the *Fix* condition was caused by inconsistent checking for password length in the key derivation function ($B_f = 30$). The second most common was failing to identify that unencrypted data was being returned from the encrypt function and encrypted data was being returned from the decrypt function ($B_f = 13$). This aligns with prior work showing that developers often assume that if their code runs and passes provided tests, then it is correct and secure [7, 19].

Participants identify a greater variety of bugs in *Read*. Conversely, in the *Read* condition, bugs that caused the code to not run or not pass the provided tests went unidentified as often as those that caused the code to fail to complete the task. The least identified in *Read* were cases where the wrong variable name was passed to or used in a function, causing the code to crash if run ($B_r = 44$). For example, about half of

participants failed to identify a bug where a Python keyword was passed to the open function in Python ($B_r = 17$), causing a crash. Participants were equally unable to identify the inconsistent check for the password length ($B_r = 16$) and returning (un)encrypted data in the encrypt and decrypt functions ($B_r = 20$). This suggests that *Read* participants, unable to run code, review all the code equally closely, resulting in identifying fewer overall but more diverse functionality issues.

4.4 Comparing security among conditions

Next, we explore how often participants who produced a functional solution were also able to produce a secure solution. We compare across all three conditions, and then discuss in detail the vulnerabilities introduced, identified, and fixed in each. Throughout this section, we use V_w, V_r, and V_f to represent the number of vulnerabilities in the *Write* condition, the *Read* condition, and the *Fix* condition, respectively. For vulnerabilities that were unique to the *Write* condition, we do not include counts for V_r and V_f.

Overall, participants in the *Write* condition produced more secure solutions than those in *Read* and *Fix*. This is perhaps unsurprising, as *Write* participants started with a blank slate, rather than starting with vulnerabilities already included.

In the *Read* and *Fix* conditions, we find little to no difference when comparing the two libraries. *Read* and *Fix* participants produced slightly more secure solutions with PyCrypto

than Cryptography.io (50% to 36% and 32% to 31% respectively), but these comparisons were not statistically significant (Table 4). Importantly, this differs from the result in the *Write* condition (Section 4.2), which (like the original paper) identified Cryptography.io as meaningfully better for security.

In all three conditions, participants were able to produce more secure solutions for the encrypt/decrypt task than the key generation and storage task, mirroring the original study.

Examining vulnerabilities in *Write*, *Read*, and *Fix*. In total, participants introduced 59 vulnerabilities in the *Write* condition, left 110 vulnerabilities unidentified in the *Read* condition, and left 176 vulnerabilities unidentified in the *Fix* condition. The large disparity in vulnerabilities between the conditions was likely due to the fact that participants in *Read* and *Fix* started with vulnerabilities in their codebase due to the study setup, while participants in the the *Write* condition started with a blank slate.

Participants assigned to use PyCrypto ($V_w = 43$, $V_r = 66$, $V_f = 88$) introduced or left unidentified more vulnerabilities than those assigned to use Cryptography.io ($V_w = 22$, $V_r = 43$, $V_f = 88$) in *Write* and *Read* conditions. This aligns with the original study and is likely due to the relative simplicity of the Cryptography.io library as well as the several secure examples within its documentation. Table 5 shows counts for vulnerability types across conditions and libraries; for explanations of vulnerabilities we inserted into the read and fix conditions, refer back to Table 1.

Participants misunderstood cryptography implementations in all conditions, but documentation weaknesses are more visible in *Write*. In every condition, the most common type of vulnerability introduced or left unidentified involved participants attempting to implement cryptography protocols but making a weak cryptography choice ($V_w = 33/59$, $V_r = 71/110$, $V_f = 110/176$). The second most common issue in all three conditions was using a fixed or static value where randomness is needed ($V_w = 16$, $V_r = 39$, $V_f = 66$).

In *Write*, both of these vulnerability types occurred overwhelmingly among participants using the PyCrypto library ($V_w = 34$ and 11 respectively) rather than Cryptography.io ($V_w = 4$ and 5 respectively), mirroring the original study. As Acar et al. note, this was likely caused by the structure of the then-current PyCrypto documentation, which made identifying and using the most secure options difficult.

Interestingly, we don't see the same pattern in the other two conditions, where vulnerabilities caused by a weak cryptography choice appear nearly equally in both libraries (Cryptography.io: $V_r = 31$, $V_f = 55$; PyCrypto: ($V_r = 40$, $V_f = 55$). Similarly, static value problems were fairly evenly distributed between the libraries in *Read* and *Fix* (Cryptography.io: $V_r = 15$, $V_f = 33$; PyCrypto ($V_r = 24$, $V_f = 33$). We hypothesize that this occurs because identifying vulnerabilities is a difficult task, regardless of the library used. This suggests that, as currently constructed, study designs using *Read* and *Fix*

would not have identified a key problem in PyCrypto that was clearly evident in the original study. We hypothesize that better documentation and simpler APIs (as found in Cryptography.io at the time) have a larger effect when writing code, but are less salient when trying to identify pre-existing bugs, two very different processes.

Focus on testing in *Fix* causes vulnerabilities to be missed. Participants correctly identified 75 out of 185 vulnerabilities in the *Read* condition but only 24 out of 200 vulnerabilities in the *Fix* condition. Our final Poisson regression model estimates that *Read* participants identified $1.65\times$ more vulnerabilities than *Fix* participants ($p < 0.001$). As with functionality, we attribute this to *Fix* participants' extreme prioritization of passing the provided tests: every single *Fix* participant started the study by running the code first, and 31 of 40 moved on immediately as soon as the code ran successfully. As noted above, this aligns with prior work regarding developers' assumptions that runnable code is correct [8, 19].

Interestingly, participants identified 8 items in the *Read* condition but only 1 item in the *Fix* condition that were non-vulnerabilities. Some of these were valid security-relevant issues outside the scope of the assigned tasks (e.g., including integrity checks for the encrypted data). Others reflected conceptual misunderstandings (e.g., two *Read* participants flagged that encryption was missing an initialization vector, but failed to notice that the code used ECB mode, which does not require an initialization vector but is highly insecure). This result also suggests that *Read* participants paid closer attention to details (even when getting some of them wrong), most likely because they could not run the code to get feedback.

Once vulnerabilities are found, *Read* and *Fix* participants face similar challenges remediating them. Not only did *Read* participants identify more vulnerabilities than *Fix* participants, they were also better at successfully remediating vulnerabilities ($V_r = 61/75$ vs. $V_f = 12/24$, respectively). However, they tended to struggle with remediating similar issues. In one notable example, 10 participants in the *Read* condition and 3 participants in the *Fix* condition noticed that the provided code used insufficient iterations in the key derivation function. The recommended value (at the time of the original study) is 10,000, but several participants ($V_r = 4$, $V_f = 1$) increased the value to only 1000, as recommended by the PyCrypto documentation. Only three of the 22 *Write* participants who used a key derivation function failed to use at least 10,000 iterations. Here, the *Read* and *Fix* study designs are able to illuminate a problem in the library documentation.

4.5 Effects on participants

In this section, we discuss the effect of the different experimental conditions on participants and response quality, including dropout rate, time to complete, and reported frustration and fun. Throughout this section we use N_r, N_w, and N_f to

Factor	Coeff	C.I.	*p*
Read	-7.3	[-22.2, 7.6]	0.341
Fix	-15.7	[-30.4, -1.1]	0.037*

Table 6: Final linear regression for completion time.

represent the number of participants in *Read*, *Write*, and the *Fix* condition, respectively.

Dropouts. Overall, only 11 of 141 participants dropped out, (N_w = 6/41, N_r = 3/54, N_f = 2/46), which is drastically different than the experience of the researchers in Acar et al. [2], who had a dropout rate of nearly 84%. This is likely due primarily to the fact that we compensated participants, but only if they completed the study, thus incentivizing them to finish. (Another contributing factor is likely that we did not include the KeyCzar and M2Crypto libraries or asymmetric encryption tasks, all of which were associated with especially high dropout rates in the original study.) Details of the number of valid and completed submissions can be seen in Table 3.

Completion time. We use the time spent on each condition as a first, crude measure of stress on participants, as participants consider time as an important factor when enrolling in secure software development studies [45]. To calculate this, we measured the time spent *actively* working on the study, i.e. excluding any breaks, by measuring the amount of time participants had our study actively open on their screen. The mean completion time for the study as a whole was 30.2 minutes (η = 19.6 minutes, σ = 32.6 minutes). Participants spent 38.2 minutes on average in the *Write* condition (η = 19.9 minutes, σ = 43.7 minutes), 30.9 minutes on average in the *Read* condition (η = 22.3 minutes, σ = 25.7 minutes), and 22.5 minutes on average in the *Fix* condition (η = 16.6 minutes, σ = 25.1 minutes). Using a linear regression, we found that participants in the *Fix* condition spent significantly less time than those in the *Write* condition ($p = 0.037$, $CI = [-30.37, -1.11]$). Details of this regression can be found in Table 6. We note that while not substantially different across conditions, we do see the fewest dropouts in *Fix*, the condition with the shortest completion time.

***Fix* reported as least frustrating, *Write* as most frustrating.** To measure study stress on participants, we asked them to self-report their frustration with the required tasks. We used an ordinal logistic regression for each task (encrypt/decrypt and key generation and storage) to understand the impact of condition/library on frustration as described in Section 3.2.

For the encrypt/decrypt task, 10 participants (29%) in the *Write* condition, 10 participants (27%) in the *Read* condition, and 6 participants (15%) in the *Fix* condition reported being frustrated (agree and strongly agree). Our model estimates that participants in the *Fix* condition were 2.71× more likely to report lower frustration than those in the *Write* condition ($p =$
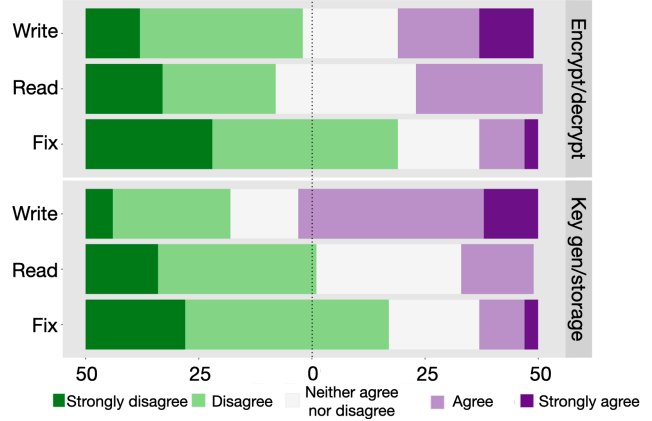
Figure 1: Reported frustration for both tasks in each condition
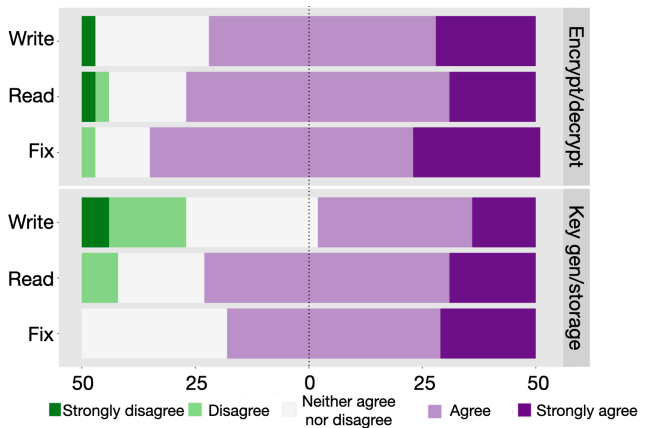
Figure 2: Reported fun for both tasks in each condition

0.02). For the key generation and storage task, 18 participants (51%) in the *Write* condition, 6 (17%) in the *Read* condition, and 4 participants (10%) in the *Fix* condition reported being frustrated (agree and strongly agree). Our model estimates that participants in the *Read* condition and the *Fix* condition were 3.7 and 6.0× more likely to report lower frustration, respectively, than those in the *Write* condition ($p = 0.004$, $p < 0.001$ respectively). We report the model and p-values in Table 7. We found no significant effect of the library on reported frustration for either task. The reported frustration in each condition for both tasks can be seen in Figure 1.

***Read* and *Fix* reported as more fun than *Write* for key generation and storage.** As another measure of study stress on participants, we asked participants to self-report whether they found the tasks fun. For the encrypt/decrypt task, 23 participants (66%) in *Write*, 28 participants (76%) in *Read*, and 34 participants (85%) in *Fix* reported having fun with the tasks (agree and strongly agree). To understand the effect of the condition and library on fun, we used the same ordinal logistic regression as described above. We found no significant effect of the condition or library on reported fun. For the

Regression	Factor	Enc/dec			Keygen		
		O.R.	C.I.	*p*	O.R.	C.I.	*p*
Frustration	Read	1.1	[0.5, 2.6]	0.834	3.7	[1.5, 9.1]	0.004*
	Fix	2.7	[1.2, 6.5]	0.023*	6.0	[2.5, 15.1]	<0.001*
	Crypto.io	0.8	[0.4, 1.7]	0.682	1.1	[0.5, 2.1]	0.887
Fun	Read	1.1	[0.4, 2.7]	0.862	2.6	[1.1, 6.4]	0.033*
	Fix	1.7	[0.7, 4.3]	0.246	2.7	[1.1, 6.5]	0.028*
	Crypto.io	1.1	[0.5, 2.2]	0.889	1.0	[0.5, 1.9]	0.909

Table 7: Ordinal logistic regression for frustration and fun.

key generation and storage task, 17 participants (49%) in the *Write* condition, 27 participants (73%) in the *Read* condition, and 36 participants (90%) in the *Fix* condition reported having fun with the tasks (agree and strongly agree). Our model estimates that participants in the *Read* and *Fix* conditions were each 2.6× and 2.7× more likely to report agreeing with having fun than those in the *Write* condition, respectively. We saw no significance of the library on reported fun. We report the model and p-values in Table 7. The reported fun in each condition for both tasks can be seen in Figure 2.

These results suggest that participants overall felt more positive about *Read* and *Fix* than *Write*, including actually enjoying *Read* and *Fix* for some tasks, with a minority of participants reporting frustration in *Fix* and a majority of participants reporting having fun in *Read* and *Fix*. A more positive participant experience has a number of potential benefits, including increased effort and better retention [45].

5 Discussion and recommendations

We now discuss how the research community can potentially apply our findings. We note that our results are exploratory in nature and that further work is likely needed to validate these findings, particularly in new contexts; we discuss this need for additional study further in Section 5.2.

5.1 When to use *Write*, *Read*, or *Fix*

Based on our experience, we make preliminary recommendations for how to use the three different study designs in future research, subject to further validation.

Use *Write* for measuring the efficacy of code writing tools. Similar to other studies that have relied on participants writing code [2, 19, 30, 41, 42, 52], our results in *Write* were able to reveal important differences between the cryptographic APIs that we tested, namely in the security of the solutions participants produced. These differences were substantially less visible in the *Read* and *Fix* conditions. We hypothesize that this may be because "simplified" cryptographic APIs are designed to prevent developers from making, rather than

identifying and fixing security mistakes. Cryptographic APIs and documentation often contain examples supporting code-writing, but these are non-exhaustive and usually don't document incorrect usage, forcing users to reference other resources when reviewing code. The *Write* method may be more appropriate when the researchers' goal is to evaluate tools aimed at *writing* secure code, such as APIs [2, 37, 50], IDE tools [18, 20, 27, 34, 41, 49], and AI-based code generation assistants [38, 39, 44], or when the goal is to categorize the security results of building a system from scratch. In particular, *Write* would be ideal for a study exploring the types of vulnerabilities that developers introduce in specific programming tasks and languages [52].

Use *Read* to measure developers' knowledge. *Read* participants identified fewer functionality bugs but more vulnerabilities than *Fix* participants, despite having no provided tests and being unable to run the code. This suggests a *Read* study design may be useful to understand the types of bugs and especially vulnerabilities developers know to look for in a given development context. This could be useful for evaluating overall security awareness and knowledge, addressing research questions about how well developers can spot problems or achieve security when they are required to pay close attention to details, providing (in some ways) an upper bound on secure development skills. For example, *Read* could be useful for understanding whether developers know they need to salt passwords to store them securely [32]. Rather than having participants build an entire system from scratch, researchers could provide participants with finished code that did (not) salt passwords, and see whether participants can spot the problem, providing a much faster approach.

Use *Fix* to measure quick fixes. Participants in the *Fix* condition found more functionality bugs than security vulnerabilities, and overall they mostly caught 'test-centric' issues, rather than deeper or less visible issues. These results echo prior work, in which participants struggled to identify new areas for testing when any tests were provided [7, 19]. This suggests *Fix* study designs may be useful for identifying vulnerabilities and bugs developers are able to quickly recognize and address using existing tests and prior knowledge. These results may therefore serve as (in some ways) a lower bound on programmers' secure development abilities.

Further, the utility of *Fix* study designs could potentially be extended to studies of secure-development tools (e.g., static analysis tools or fuzzers) that automatically flag certain types of vulnerabilities for developers' attention or of how participants select and evaluate AI-generated code suggestions, as these studies highlight how the evaluation and interaction is performed rather than just the identification. We found that once issues are identified, *Read* and *Fix* participants had roughly similar success in addressing them, so this study design may be useful for examining how effectively developers can understand and address vulnerabilities these tools iden-

tify.

Use *Read* and *Fix* to minimize time, frustration. Compared with the the *Write* condition, participants who had to *Read* or *Fix* code spent, on average, less time completing the study and found their tasks less frustrating and more fun. Prior research has found that industry developers—a desirable but hard-to-reach demographic for secure development studies—prioritize factors such as study duration and low effort [45]. Therefore, when appropriate for the research questions, *Read* and *Fix* methods may offer reasonable trade-offs to researchers who are concerned about recruiting enough target participants for studies.

5.2 How to design *Write*, *Read*, *Fix* studies

Since the *Read* and *Fix* methods for secure development studies are relatively novel, there are several design considerations future researchers should take into account. In this preliminary study, we were only able to explore a few points in this potential design space; we hope that going forward, other researchers will explore different trade-offs and design choices, to better characterize the pros and cons of different study designs in the broader secure-development context.

Improving *Write* studies. We suspect that there are limits to how enjoyable write studies can be designed to be. Our low-dropout experience, compared to Acar et al. [2], suggests compensation and clear study expectations help. Echoing prior work, our results also suggest that fun and time spent matters, so more interesting and shorter tasks may help [45].

Utilizing bonus payments. While bonus payments have not previously been widely used, we added them after observing some participants skipping all tasks to receive compensation. We specifically left the framing of the bonus vague so as to not sway participants too heavily to only address a certain number of bugs. The addition of the bonus improved the validity of our results, as it reduced the number of low-effort submissions. However, it is possible that it focused participants' attention on functionality bugs, as they were motivated to receive the bonus, possibly causing them to not look for security vulnerabilities. Researchers should consider the inclusion and framing of a bonus carefully. While it helps promote participant retention and quality, it may also sway participants in the *Fix* condition to focus solely on passing the tests, rather than taking a more holistic approach.

Inserting realistic vulnerabilities. In order to conduct a study based on reading or fixing code, researchers must insert appropriately realistic vulnerabilities (and potentially develop predefined tests or tools that can flag them). For this early-stage, exploratory work, we derived these vulnerabilities from real participant examples observed in a prior code-writing study. Of course, researchers who are using *Read* or *Fix* instead of a *Write* study are unlikely to have this kind of prior

data available. We suggest instead identifying realistic security defects from known vulnerability listings (e.g., CVE lists) or using examples taken from open-source software or from programming sites like Stack Overflow. It may also be possible to derive defects from interview or survey studies that reveal developers' misconceptions and mental models. These approaches, however, may need to be validated with further studies of experimental methods.

Enhancing ecological validity. For this early-stage study comparing methods, we prioritized internal validity (straightforward comparisons between conditions) as well as replicating prior work (for contextualizing our results). This led to specific design choices, such as restricting participants to specific libraries, avoiding external documentation, using older versions of software and documentation, and preventing *Read* participants from running code. We believe these choices made sense for this study, but they did reduce ecological validity, as real-world code review processes obviously lack many or all of these restrictions. Researchers considering employing *Read* and *Fix* should revisit these trade-offs for their own research questions and constraints; for example, in some studies it may be useful to more closely match the flow of real-world code review processes such as those used in public repositories like GitHub.

One of the main reasons *Fix* underperformed was because participants focused primarily on passing the pre-written tests; since these are not an inherent feature of *Fix*, future work should investigate whether removing them would increase the overall effectiveness of this condition. As with other exploratory changes, versions of *Read* and *Fix* that prioritize ecological validity should be validated with additional methodological studies.

Exploring *Read* and *Fix* for other secure-development domains and questions. We designed our study around a specific secure-development research question: the relative effectiveness of cryptographic APIs. Other security domains may suffer from different types of vulnerabilities and bugs, and other types of research questions may exhibit different outcomes from different study designs. Further work is needed to explore how *Read* and *Fix* would function in these different contexts. Perhaps other novel methods, in addition to the three study designs we considered, can also be developed to address these contexts.

6 Acknowledgments

References

[1] Upwork. https://www.upwork.com/.

[2] Yasemin Acar, Michael Backes, Sascha Fahl, Simson Garfinkel, Doowon Kim, Michelle L Mazurek, and Christian Stransky. Comparing the Usability of Cryptographic APIs. In *IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017.

[3] Yasemin Acar, Michael Backes, Sascha Fahl, Doowon Kim, Michelle L Mazurek, and Christian Stransky. You Get Where You're Looking For: The Impact of Information Sources on Code Security. In *IEEE Symposium on Security and Privacy (SP)*. IEEE, 2016.

[4] Yasemin Acar, Sascha Fahl, and Michelle L Mazurek. You are Not Your Developer, Either: A Research Agenda for Usable Security and Privacy Research Beyond End Users. In *IEEE Cybersecurity Development (SecDev)*. IEEE, 2016.

[5] Yasemin Acar, Christian Stransky, Dominik Wermke, Michelle L Mazurek, and Sascha Fahl. Security Developer Studies with GitHub Users: Exploring a Convenience Sample. In *Symposium on Usable Privacy and Security (SOUPS)*, 2017.

[6] Noura Alomar, Primal Wijesekera, Edward Qiu, and Serge Egelman. "You've Got Your Nice List of Bugs, Now What?" Vulnerability Discovery and Management Processes in the Wild. In *Symposium on Usable Privacy and Security (SOUPS)*, 2020.

[7] Hala Assal and Sonia Chiasson. Security in the Software Development Lifecycle. In *Symposium on Usable Privacy and Security (SOUPS)*, 2018.

[8] Wei Bai, Omer Akgul, and Michelle L Mazurek. A Qualitative Investigation of Insecure Code Propagation from Online Forums. In *IEEE Cybersecurity Development (SecDev)*. IEEE, 2019.

[9] Sebastian Baltes and Stephan Diehl. Worse Than Spam: Issues in Sampling Software Developers. In *ACM/IEEE international symposium on empirical software engineering and measurement*, 2016.

[10] Gunnar R Bergersen, Dag IK Sjøberg, and Tore Dybå. Construction and validation of an instrument for measuring programming skill. *IEEE Transactions on Software Engineering*, 40(12), 2014.

[11] Veroniek Binkhorst, Tobias Fiebig, Katharina Krombholz, Wolter Pieters, and Katsiaryna Labunets. Security at the End of the Tunnel: The Anatomy of VPN Mental Models Among Experts and Non-Experts in a Corporate Context. In *USENIX Security Symposium (USENIX Security)*, 2022.

[12] Yung-Yu Chang, Pavol Zavarsky, Ron Ruhl, and Dale Lindskog. Trend Analysis of the CVE for Software Vulnerability Management. In *IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. IEEE, 2011.

[13] Anastasia Danilova, Stefan Horstmann, Matthew Smith, and Alena Naiakshina. Testing Time Limits in Screener Questions for Online Surveys with Programmers. In *IEEE International Conference on Software Engineering (ICSE)*, 2022.

[14] Anastasia Danilova, Alena Naiakshina, Johanna Deuter, and Matthew Smith. Replication: On the Ecological Validity of Online Security Developer Studies: Exploring Deception in a Password-Storage Study with Freelancers. 2020.

[15] Anastasia Danilova, Alena Naiakshina, Stefan Horstmann, and Matthew Smith. Do you really code? Designing and Evaluating Screening Questions for Online Surveys with Programmers. In *International Conference on Software Engineering (ICSE)*. IEEE, 2021.

[16] Anastasia Danilova, Alena Naiakshina, Anna Rasgauski, and Matthew Smith. Code Reviewing as Methodology for Online Security Studies with Developers-A Case Study with Freelancers on Password Storage. In *Symposium on Usable Privacy and Security (SOUPS)*, 2021.

[17] Janet Feigenspan, Christian Kästner, Jörg Liebig, Sven Apel, and Stefan Hanenberg. Measuring programming experience. In *IEEE international conference on program comprehension (ICPC)*. IEEE, 2012.

[18] Matthew Finifter and David Wagner. Exploring the Relationship Between Web Application Development Tools and Security. 2011.

[19] Kelsey R Fulton, Daniel Votipka, Desiree Abrokwa, Michelle L Mazurek, Michael Hicks, and James Parker. Understanding the How and the Why: Exploring Secure Development Practices through a Course Competition. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2022.

[20] Peter Leo Gorski, Luigi Lo Iacono, Dominik Wermke, Christian Stransky, Sebastian Möller, Yasemin Acar, and Sascha Fahl. Developers Deserve Security Warnings, Too: On the Effect of Integrated Security Advice on Cryptographic API Misuse. In *Symposium on Usable Privacy and Security (SOUPS)*, 2018.

[21] Matthew Green and Matthew Smith. Developers are Not the Enemy!: The Need for Usable Security APIs. *IEEE Security & Privacy*, 14(5), 2016.

[22] Andrew F Hayes and Klaus Krippendorff. Answering the Call for a Standard Reliability Measure for Coding Data. *Communication methods and measures*, 1(1).

[23] Mohammadreza Hazhirpasand, Oscar Nierstrasz, Mohammadhossein Shabani, and Mohammad Ghafari. Hurdles for Developers in Cryptography. 2021.

[24] Nicolas Huaman, Alexander Krause, Dominik Wermke, Jan H. Klemmer, Christian Stransky, Yasemin Acar, and Sascha Fahl. If You Can't Get Them to the Lab: Evaluating a Virtual Study Environment with Security Information Workers. In *Symposium on Usable Privacy and Security (SOUPS)*, 2022.

[25] Harjot Kaur, Sabrina Amft, Daniel Votipka, Yasemin Acar, and Sascha Fahl. Where to Recruit for Security Development Studies: Comparing Six Software Developer Samples. 2022.

[26] Joe Lewis. Nerds. https://github.com/SP2-MC2/Developer-Observatory, 2021.

[27] Tianshi Li, Yuvraj Agarwal, and Jason I Hong. Coconut: An IDE Plugin for Developing Privacy-Friendly Apps. *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(4), 2018.

[28] Abraham H Mhaidli, Yixin Zou, and Florian Schaub. "We Can't Live Without Them!" App Developers' Adoption of Ad Networks and Their Considerations of Consumer Risks. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, pages 225–244, 2019.

[29] Mitre. CVE. https://cve.mitre.org/, 2020.

[30] Alena Naiakshina, Anastasia Danilova, Eva Gerlitz, and Matthew Smith. On Conducting Security Developer Studies with CS Students: Examining a Password-Storage Study with CS Students, Freelancers, and Company Developers. In *CHI Conference on Human Factors in Computing Systems*, 2020.

[31] Alena Naiakshina, Anastasia Danilova, Eva Gerlitz, Emanuel Von Zezschwitz, and Matthew Smith. "If you want, I can store the encrypted password": A Password-Storage Field Study with Freelance Developers. In *CHI Conference on Human Factors in Computing Systems*, 2019.

[32] Alena Naiakshina, Anastasia Danilova, Christian Tiefenau, Marco Herzog, Sergej Dechand, and Matthew Smith. Why Do Developers get Password Storage Wrong? A Qualitative Usability Study. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2017.

[33] Alena Naiakshina, Anastasia Danilova, Christian Tiefenau, and Matthew Smith. Deception Task Design in Developer Password Studies: Exploring a Student Sample. In *Symposium on Usable Privacy and Security (SOUPS)*, 2018.

[34] Duc Cuong Nguyen, Dominik Wermke, Yasemin Acar, Michael Backes, Charles Weir, and Sascha Fahl. A Stitch in Time: Supporting Android Developers in Writing Secure Code. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2017.

[35] NIST. National Vulnerability Database. https://nvd.nist.gov/general, 2020.

[36] Anna-Marie Ortloff, Christian Tiefenau, and Matthew Smith. SoK: I Have the (Developer) Power! Sample Size Estimation for Fisher's Exact, Chi-Squared, McNemar's, Wilcoxon Rank-Sum, Wilcoxon Signed-Rank and t-tests in Developer-Centered Usable Security. In *Symposium on Usable Privacy and Security (SOUPS)*, 2023.

[37] Nikhil Patnaik, Joseph Hallett, and Awais Rashid. Usability Smells: An Analysis of Developers' Struggle With Crypto Libraries. 2019.

[38] Hammond Pearce, Baleegh Ahmad, Benjamin Tan, Brendan Dolan-Gavitt, and Ramesh Karri. Asleep at the Keyboard? Assessing the Security of GitHub Copilot's Code Contributions. In *IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022.

[39] Neil Perry, Megha Srivastava, Deepak Kumar, and Dan Boneh. Do Users Write More Insecure Code with AI Assistants? In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2023.

[40] Olgierd Pieczul, Simon Foley, and Mary Ellen Zurko. Developer-centered security and the symmetry of ignorance. In *New Security Paradigms Workshop*, 2017.

[41] Stephan Plöger, Mischa Meier, and Matthew Smith. A Qualitative Usability Evaluation of the Clang Static Analyzer and libFuzzer with CS Students and CTF Players. In *Symposium on Usable Privacy and Security (SOUPS)*, 2021.

[42] Andrew Ruef, Michael Hicks, James Parker, Dave Levin, Michelle L Mazurek, and Piotr Mardziel. Build It, Break It, Fix It: Contesting Secure Development. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2016.

[43] Iflaah Salman, Ayse Tosun Misirli, and Natalia Juristo. Are Students Representatives of Professionals in Software Engineering Experiments? In *IEEE International Conference on Software Engineering (ICSE)*, 2015.

[44] Gustavo Sandoval, Hammond Pearce, Teo Nys, Ramesh Karri, Siddharth Garg, and Brendan Dolan-Gavitt. Lost at C: A User Study on the Security Implications of Large Language Model Code Assistants. In *USENIX Security Symposium (USENIX Security)*, 2023.

[45] Raphael Serafini, Marco Gutfleisch, Stefan Albert Horstmann, and Alena Naiakshina. On the Recruitment of Company Developers for Security Studies: Results from a Qualitative Interview Study. In *Symposium on Usable Privacy and Security (SOUPS)*, 2023.

[46] Matthew Smith. Usable Security—The Source Awakens. USENIX Association, 2016.

[47] Christian Stransky, Yasemin Acar, Duc Cuong Nguyen, Dominik Wermke, Doowon Kim, Elissa M. Redmiles, Michael Backes, Simson Garfinkel, Michelle L. Mazurek, and Sascha Fahl. Lessons Learned from Using an Online Platform to Conduct Large-Scale, Online Controlled Security Experiments with Software Developers. In *USENIX Workshop on Cyber Security Experimentation and Test (CSET)*. USENIX Association, 2017.

[48] Mohammad Tahaei and Kami Vaniea. Recruiting Participants With Programming Skills: A Comparison of Four Crowdsourcing Platforms and a CS Student Mailing List. In *CHI Conference on Human Factors in Computing Systems*, 2022.

[49] Mohammad Tahaei, Kami Vaniea, Konstantin Beznosov, and Maria K Wolters. Security Notifications in Static Analysis Tools: Developers' Attitudes, Comprehension, and Ability to Act on Them. In *CHI Conference on Human Factors in Computing Systems*, 2021.

[50] Christian Tiefenau, Emanuel von Zezschwitz, Maximilian Häring, Katharina Krombholz, and Matthew Smith. A Usability Evaluation of Let's Encrypt and Certbot: Usable Security Done Right. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 1971–1988, 2019.

[51] Daniel Votipka, Desiree Abrokwa, and Michelle L Mazurek. Building and Validating a Scale for Secure Software Development Self-Efficacy. In *CHI Conference on Human Factors in Computing Systems*, 2020.

[52] Daniel Votipka, Kelsey R Fulton, James Parker, Matthew Hou, Michelle L Mazurek, and Michael Hicks. Understanding security mistakes developers make: Qualitative analysis from Build It, Break It, Fix It. In *USENIX Security Symposium (USENIX Security)*, 2020.

[53] Aiko Yamashita and Leon Moonen. Surveying developer knowledge and interest in code smells through online freelance marketplaces. In *International Workshop on User Evaluations for Software Engineering Researchers (USER)*. IEEE, 2013.

A Survey instruments

A.1 Final survey

A.1.1 Encrypt/decrypt task specific questions

1. Recall your experiences with the task where you were expected to encrypt/decrypt data (**encryption/decryption task**).
2. I completed the **encryption/decryption task** correctly.
   - I am not confident.
   - I am slightly confident.
   - I am somewhat confident.
   - I am moderately confident.
   - I am absolutely confident.
3. I completed the **encryption/decryption task** securely.
   - I am not confident.
   - I am slightly confident.
   - I am somewhat confident.
   - I am moderately confident.
   - I am absolutely confident.
4. The documentation was helpful in completing the **encryption/decryption task**.
   - Strongly agree
   - Agree
   - Neither agree nor disagree
   - Disagree
   - Strongly disagree
5. Completing the **encryption/decryption task** was frustrating.
   - Strongly agree
   - Agree
   - Neither agree nor disagree
   - Disagree
   - Strongly disagree
6. Completing the **encryption/decryption task** was fun.
   - Strongly agree
   - Agree
   - Neither agree nor disagree
   - Disagree
   - Strongly disagree
7. Completing the **encryption/decryption task** was tedious.
   - Strongly agree
   - Agree
   - Neither agree nor disagree
   - Disagree
   - Strongly disagree
8. Completing the **encryption/decryption task** was challenging.

- Strongly agree
- Agree
- Neither agree nor disagree
- Disagree
- Strongly disagree

9. What parts of the **encryption/decryption task** were easy?
    - Text box
10. What parts of the **encryption/decryption task** were difficult?
    - Text box

A.1.2 Keygen task specific questions

1. Recall your experiences with the task where you were expected to expected to generate an encryption key and store it securely (**key generation and storage task**).
2. I completed the **key generation and storage task** correctly.
    - I am not confident.
    - I am slightly confident.
    - I am somewhat confident.
    - I am moderately confident.
    - I am absolutely confident.
3. I completed the **key generation and storage task** securely.
    - I am not confident.
    - I am slightly confident.
    - I am somewhat confident.
    - I am moderately confident.
    - I am absolutely confident.
4. The documentation was helpful in completing the **key generation and storage task**.
    - Strongly agree
    - Agree
    - Neither agree nor disagree
    - Disagree
    - Strongly disagree
5. Completing the **key generation and storage task** was frustrating.
    - Strongly agree
    - Agree
    - Neither agree nor disagree
    - Disagree
    - Strongly disagree
6. Completing the **key generation and storage task** was fun.
    - Strongly agree
    - Agree
    - Neither agree nor disagree
    - Disagree
    - Strongly disagree
7. Completing the **key generation and storage task** was tedious.

- Strongly agree
- Agree
- Neither agree nor disagree
- Disagree
- Strongly disagree

8. Completing the **key generation and storage task** was challenging.
    - Strongly agree
    - Agree
    - Neither agree nor disagree
    - Disagree
    - Strongly disagree
9. What parts of the **key generation and storage task** were easy?
    - Text box
10. What parts of the **key generation and storage task** were difficult?
    - Text box

A.1.3 Study specific questions

1. Are you aware of a specific library or other resource you would have preferred to use to generate functional and secure code? If yes, please list them.
    - Yes [Text box]
    - No
2. Have you used or seen this assigned library before?
    - I have used the assigned library before. (e.g. worked on a project with assigned library)
    - I have seen code from the assigned library but not used it myself. (e.g. worked on a project with the library but someone else wrote the code)
    - I have neither used nor seen the assigned library before.
    - I don't know
3. Have you written or seen code for tasks similar to the assigned tasks before?
    - I have written similar code. (e.g. worked on a project that included a similar task)
    - I have seen similar code but have not written it myself. (e.g. worked on a project that included a similar task but someone else wrote the code)
    - I have never written nor seen code for similar tasks.
    - I don't know

A.1.4 System Usability Scale

1. We asked you to use an assigned library. To what extent do you agree with each of the following statements in reference to your assigned library and it's documentation: (Strongly agree, Agree, Neither agree nor disagree, Disagree, Strongly disagree)
    - I think that I would like to use this library frequently.

- I found this library unnecessarily complex.
- I thought this library was easy to use.
- I think that I would need the support of a technical person to be able to use this library.
- I found the various functions in this library were well integrated.
- I found this library fun to use. Regardless of what you felt please select strongly agree.
- I thought there was too much inconsistency in this library.
- I would imagine that most people would learn to use this library very quickly.
- I found this library very cumbersome to use.
- I felt very confident using this library.
- I needed to learn a lot of things before I could get going with this library.

A.1.5 Acar Usability Scale

1. We asked you to use an assigned library. To what extent do you agree with each of the following statements in reference to your assigned library and it's documentation: (Strongly agree, Agree, Neither agree nor disagree, Disagree, Strongly disagree)
    - I had to understand how most of the assigned library works in order to complete the tasks.
    - It would be easy and require only small changes to change parameters or configuration later without breaking my code.
    - After doing these tasks, I think I have a good understanding of the assigned library overall.
    - I only had to read a little of the documentation for the assigned library to understand the concepts that I needed for these tasks.
    - The names of classes and methods in the assigned library corresponded well to the functions they provided.
    - It was straightforward and easy to implement the given tasks using the assigned library.
    - When I accessed the assigned library documentation, it was easy to find useful help.
    - In the documentation, I found helpful explanations.
    - In the documentation, I found helpful code examples.
    - When I made a mistake, I got a meaningful error message/exception.
    - Using the information from the error message/exception, it was easy to fix my mistake.
    - Using the information from the error message/exception, it was hard to fix. Please select strongly disagree.

A.1.6 SSD-SES

1. During this portion of the survey, you will be shown hypothetical software development tasks. Please rate your level of confidence in completing the following software development tasks. (I am not confident, I am slightly confident, I am somewhat confident, I am moderately confident, I am absolutely confident, I do not understand the question)
    - I can perform a threat risk analysis (e.g. likelihood of vulnerability, impact of exploitation, etc.)
    - I can identify potential security threats to the system.
    - I can identify common attack techniques used by attackers.
    - I can identify potential attack vectors in the environment the system interacts with (e.g., hardware, libraries, etc.).
    - I can identify common vulnerabilities of a programming language.
    - I can design software to quarantine an attacker if a vulnerability is exploited.
    - I can mimic potential threats to the system.
    - I can evaluate security controls on the system's interfaces/interactions with other software systems.
    - I can evaluate security controls on the system's interfaces/interactions with other hardware systems.
    - I can communicate security assumptions and requirements to other developers on the team to ensure vulnerabilities are not introduced due to misunderstandings.
    - I can communicate system details with other developers to ensure a thorough security review of the code.
    - I can discuss lessons learned from internal and external security incidents to ensure all development team members are aware of potential threats.
    - I can effectively communicate identified security issues and the cost/risk trade-off associated with deciding whether or not to fix the problem to organization leadership.
    - I can communicate functionality needs to security experts to get recommendations for secure solutions (e.g., secure libraries, design patterns, and platforms).
    - I know the appropriate point of contact/response team in my organization to contact if a vulnerability in production code is identified.
    - I can perform security assessments. Regardless of your actual answer, please select I am absolutely confident.

A.1.7 General technical background

1. Including education, how long have you been programming? (In years)
   - Text box
2. Including education, how long have you been programming in Python? (In years)
   - Text box
3. Are you currently employed in a role that requires programming?
   - Yes
   - No
   - Maybe
4. (If yes or maybe to above) Is writing code in Python part of your primary job?
   - Yes
   - No
   - Maybe
5. (If yes or maybe to above) Not including education, how long have you been programming professionally? (In years)
   - Text box
6. (If yes or maybe to above) Not including education, how long have you been programming in Python professionally? (In years)
   - Text box
7. (If yes or maybe to above) Which of the following job roles describe you? (Please select all that apply)
   - Developer
   - Administrator
   - DevOps Engineer
   - Academic researcher/Scientist
   - Data science/Machine learning specialist
   - Educator
   - Engineer
   - Manager/Team lead
   - None
   - Other [Text box]
8. How did you learn to code? (Please select all that apply)
   - Self-taught
   - Online class
   - College/University
   - On-the-job training
   - Professional certification program
   - Coding bootcamp
   - I did not learn to code
   - Other [Text box]
9. How do you rate your knowledge of software security?
   - Very high
   - Above average
   - Average
   - Below average
   - Very low
10. Which of the following statements describe the secure programming training that you have received? (Please select all that apply)
    - I received secure programming training through an event organized by my employer
    - I learned secure programming concepts while working
    - I received secure programming training at school/college/university
    - I received secure programming training at a workshop/seminar
    - I received secure programming training with online courses
    - I am self-taught
    - I have never received secure programming training
11. How many total years of experience do you have in computer security? (Experience includes years at work or studying in a security-related field)
    - Text box

A.1.8 Demographics

1. Please select the gender with which you most closely identify:
   - Man
   - Woman
   - Non-binary
   - Another gender/prefer to self-describe [Text box]
   - Prefer not to answer
2. What is your age in years?
   - Text box
3. Please specify your ethnicity. (Please select all that apply)
   - White
   - Hispanic or Latino
   - Black or African American
   - American Indian or Alaskan Native
   - Asian
   - Native Hawaiian or Pacific Islander
   - Prefer to self-describe [Text box]
   - Prefer not answer
4. Please select your highest completed education level.
   - Some high school
   - High school diploma/GED
   - Some college, no degree
   - Associate's degree
   - Bachelor's degree
   - Master's degree
   - Doctoral degree
   - Prefer not to answer
5. (If college or above) What was your primary field of study?
   - Computer science
   - IT security/Cyber security
   - Other engineering disciplines

- Never declared a major
- Other [Text box]
6. What is your country of residence?
   - Text box

A.2 Screening survey

A.2.1 General background

1. How long have you been programming?
   - Less than 1 year
   - 1 - 2 years
   - 2 - 5 years
   - More than 5 years
2. Are you currently a student?
   - Yes
   - No
3. (If yes to above) Are you currently majoring in something that requires programming?
   - Yes
   - Maybe
   - No
4. (If yes or maybe to above) What is your major?
   - Text box
5. Are you currently employed in a role that requires programming?
   - Yes
   - Maybe
   - No
6. (If yes or maybe to above) What is your occupation?
   - Text box
7. Please rate your proficiency with the following languages:
   - Java (Extremely proficient, Moderately proficient, Somewhat proficient, Not at all proficient, I am not familiar with this programming language)
   - C (Extremely proficient, Moderately proficient, Somewhat proficient, Not at all proficient, I am not familiar with this programming language)
   - C++ (Extremely proficient, Moderately proficient, Somewhat proficient, Not at all proficient, I am not familiar with this programming language)
   - Python (Extremely proficient, Moderately proficient, Somewhat proficient, Not at all proficient, I am not familiar with this programming language)
   - Rust (Extremely proficient, Moderately proficient, Somewhat proficient, Not at all proficient, I am not familiar with this programming language)
   - Ruby (Extremely proficient, Moderately proficient, Somewhat proficient, Not at all proficient, I am not familiar with this programming language)
   - Javascript (Extremely proficient, Moderately proficient, Somewhat proficient, Not at all proficient, I am not familiar with this programming language)

- OCaml (Extremely proficient, Moderately proficient, Somewhat proficient, Not at all proficient, I am not familiar with this programming language)

A.2.2 Screener from Danilova et al. [15]

1. Which of the following websites do you most frequently use as an aid when programming?
   - Wikipedia
   - LinkedIn
   - StackOverflow
   - Memory Alpha
   - I have not used any of the websites above for programming.
   - I don't program
2. Choose the answer that best fits the description of a compiler's function.
   - Refactoring code
   - Connecting to the network
   - Aggregating user data
   - Translating code into executable instructions
   - Collecting user data
   - I don't know
3. Choose the answer that best fits the definition of a recursive function.
   - A function that runs for infinite time
   - A function that does not have a return value
   - A function that can be called from other functions
   - A function that calls itself
   - A function that does not require an input
   - A function that interprets cursive handwriting
   - I don't know
4. Which of these values could be assigned to a variable with the type Boolean?
   - Small
   - Solid
   - Quadratic
   - Red
   - True
   - I don't know
5. Answer the next two questions using the following snippet:

```python
def func(example):
    x = len(example)
    out = ""
    for i in range(x):
        out = out + example[x - i - 1]
    return out

print(func("hello_world"))
```

6. Referring to the above code snippet, what is the parameter of the function?
   - out
   - example
   - for i in range(x)

- Outputting a string
- x = len(example)
- I don't know
7. Please select the returned value of the pseudocode above:
    - hello world
    - hello world 10
    - dlrow olleh
    - world hello
    - HELLO WORLD
    - hello world hello world hello world hello world
    - I don't know

Evaluating Privacy Perceptions, Experience, and Behavior of Software Development Teams

Maxwell Prybylo
University of Maine

Sara Haghighi
University of Maine

Sai Teja Peddinti
Google

Sepideh Ghanavati
University of Maine

Abstract

With the increase in the number of privacy regulations, small development teams are forced to make privacy decisions on their own. In this paper, we conduct a mixed-method survey study, including statistical and qualitative analysis, to evaluate the privacy perceptions, practices, and knowledge of members involved in various phases of the Software Development Life Cycle (SDLC). Our survey includes 362 participants from 23 countries, encompassing roles such as product managers, developers, and testers. Our results show diverse definitions of privacy across SDLC roles, emphasizing the need for a holistic privacy approach throughout SDLC. We find that software teams, regardless of their region, are less familiar with privacy concepts (such as anonymization), relying on self-teaching and forums. Most participants are more familiar with GDPR and HIPAA than other regulations, with multi-jurisdictional compliance being their primary concern. Our results advocate the need for role-dependent solutions to address the privacy challenges, and we highlight research directions and educational takeaways to help improve privacy-aware SDLC.

1   Introduction

With the vast increase in privacy violations in the US and around the world [95], many countries have adopted new privacy regulations [66], such as the European General Data Protection Regulation (GDPR) [27]. With these new regulations, developers are under increased scrutiny while implementing privacy engineering solutions throughout the Software Development Life Cycle (SDLC) or face financial penalties. Many

mobile apps are initially developed by a small team of independent developers with limited privacy expertise or access to legal/policy resources to make privacy decisions [6, 7, 63]. Research shows that this lack of access to privacy expertise leads to challenges in creating concise, accurate and consistent privacy policies [11, 13, 52, 65, 72, 77, 78, 96, 99], implementing privacy concepts throughout the SDLC - from early analysis to testing [26, 40, 63, 86], and distinguishing between privacy and security approaches, tools and regulations [7, 9, 21, 36, 40, 81, 82, 89].

In recent years, several approaches, including Privacy by Design (PbD), have been introduced to help developers incorporate privacy rules throughout the SDLC [17, 20, 31–33, 44, 45, 53, 55, 65, 80, 99]. However, few works examined the implementation of these solutions from the developers' perspective and their impact on privacy practices. Most studies focus on only a limited group of developers and overlook the broader SDLC roles and the unique challenges faced by each role (e.g., product manager when defining privacy requirements or the QAs when identifying privacy leaks) [21, 48, 51, 61, 91]. They also do not examine how factors such as legal expertise, regulations, and regional differences influence software teams' privacy perceptions and practices.

In this paper, we conduct a large mixed-method survey study on Prolific with 189 participants located in the US and 173 participants located in 22 other countries (in total 362), who are involved in various roles in the SDLC – including administrators (e.g., scrum masters, product managers), development and Quality Assurance (QA) teams, and information security/privacy experts. The non-US participants are located in EU+UK (132), South Africa (21), Mexico (15), Canada (3), and South America (2). Our goal is to identify the current state of privacy comprehension, practices, and behaviors in various SDLC roles, and the privacy gaps that have yet to be addressed. Our survey comprises of three parts: pre-screening questions (e.g., describing their product/customers), generic questions regarding participants' demographics (education, role, company size, etc.); and role-specific questions to examine their perceptions, experiences, and behaviors. We combine

the participants' demographics (e.g., location data) with the role-specific responses to help determine:

– **RQ1:** Are there any differences in privacy perceptions among various roles, locations, and other demographics?
– **RQ2:** Does access to privacy experts (e.g., a Chief Privacy Officer - CPO) impact privacy perceptions and practices?
– **RQ3:** How do privacy practices and experiences vary according to SDLC roles, locations, and other demographics?
– **RQ4:** What is the degree of familiarity of different roles regarding privacy concepts, approaches, tools, and regulations?

To the best of our knowledge, this is the first study to conduct such a holistic evaluation based on the roles in the SDLC.

Our results show that participants have diverse perceptions/definitions of privacy, showcasing the need for a refined approach to privacy in SDLC. Scrum masters, product managers, and information security/privacy experts define privacy more in terms of limited disclosure, while developers and QAs perceive privacy as control over personal information. Our study finds a lack of adoption of most PbD strategies and other privacy techniques, such as Privacy Enhancing Technologies (PETs) and Privacy Impact Assessment (PIA), in SDLC. Most QA members rely on legal/privacy experts to protect users' data, and they lack privacy knowledge and expertise. Members of software teams are generally self-taught regarding privacy concepts, and most are not familiar with regulations that exist in the US, such as the California Consumer Privacy Act (CCPA) [35] and the Children's Online Privacy Protection Rule (COPPA) [29]. We also find that software teams face challenges in both understanding and adhering to privacy regulations, especially across multiple jurisdictions. These findings highlight the need for more privacy-focused education and training. Comparing regional-specific trends regarding the use of PETs, the creation of PIAs, or the presence of a CPO, we did not observe any differences among our participants, regardless of their location. This shows that privacy practices are primarily determined by the culture of the organization and are not influenced by various regulations across regions [5]. Our results highlight the challenges faced in various SDLC roles and advocate the need for role-dependent solutions to address them. Based on these findings, we outline research directions and educational takeaways to help improve privacy-aware SDLC.

2 Related Work

Understanding developers' privacy expertise and concerns has been explored in research through user studies with developers and analysis of developers' forums [21, 48, 51, 61, 91]. Tahaei et al. [85] and Horstmann et al. [48] conducted interviews with developers and privacy experts and identified factors such as poor privacy culture, tensions between privacy and other business rules, lack of proper communication between privacy experts and developers, lack of standardized privacy tools, and mismatch between the technical expertise

of developers and privacy experts that impact how developers implement privacy. They also emphasize the role of privacy champions to minimize such barriers. In 2014 (i.e., pre-GDPR and CCPA), Balebako et al. [7] examined how app developers make privacy and security decisions and revealed that smaller companies exhibit fewer positive privacy and security behaviors. Their research emphasizes the need for simplified, cost-effective privacy tools such as privacy checklists, especially for small firms. Other studies [3, 48, 51, 57] with practitioners and developers highlight that while regulations impact practitioners' behaviors and corporate cultures, the developers and practitioners mostly rely on app markets to spot privacy issues, and they struggle with implementing and maintaining privacy labels, as well as leveraging third-party tools to maintain compliance.

The analysis of Stack Overflow (SO) posts shows that developers frequently query regarding PbD, compliance, and confidentiality [84, 91]. Delile et al. [24] compared privacy questions on SO with responses generated by ChatGPT to identify whether ChatGPT could be used as an alternative tool. Their results indicate that, in ∼30% of cases SO is more accurate than ChatGPT. Li et al. [59] and Parsons et al. [70] studied posts on several Reddit forums and identified that most discussions on personal data usage occur in response to external events such as Android OS changes or privacy laws.

These studies pinpoint developers' challenges in correctly implementing privacy requirements and maintaining compliance. Our work complements these efforts; however, it is the first study to assess privacy perceptions, practices, and knowledge of members of software teams involved in various roles in SDLC through a large-scale mixed-method approach. Prior work focused only on developers (i.e., programmers) in the US and a few countries, whereas we studied members from various SDLC roles (including product managers, QA, etc.) spanning 23 countries. In our work, we investigate how factors such as organizational aspects (e.g., the presence of a CPO) and participants' demographics (e.g., role, education, and location) impact privacy perceptions, experience, and behaviors of software teams. We also explore how frequently developers use online forums for privacy-related queries.

3 Study Design

In this paper, we aim to understand how members of software teams in small, medium-sized, and large companies (i.e., with <20, 21–100, and 100+ employees), implement privacy in their software applications and examine their level of privacy comprehension, expertise, practices, and behaviors based on various demographics (such as roles, location, education level, etc.). For this purpose, we first conducted a pilot study to evaluate our survey design and then a large-scale study with members of software teams in 23 countries. Our pilot study was completed in January 2023, while our large-scale study was done between February–April 2023.

Table 1: Breakdown of Participants Roles

Role	Count
AD: Admin., Product Manager, Scrum Master	70
SD: Software Designer, Architect, Developer	198
QA: Software Tester, Quality Assurance Eng.	40
ISec: Information Security/Privacy Expert	54
Total	**362**

3.1 Survey Tools

Survey Creation We utilized *Qualtrics* for survey creation, a platform supported by our university. Using Qualtrics, we customized our survey to individualize questions based on the participants' role (Q9) as defined in Table 1. For example, we asked developers about familiarity with PbD (Q39) and their use of forums such as Reddit (Q32), while information security members were asked about the management of access control, encryption algorithms, and certificates (Q60-Q62).

Survey Platform Selection We conducted the large-scale survey using Qualtrics integration on the Prolific [68, 90] platform, since it provides a higher pay rate and allows selecting from a more specific pool of participants with basic programming knowledge, in our case - software teams.

Tahaei et al. and Kaur et al. [56, 90] recommend using Prolific and MTurk for large-scale surveys. Although pre-screening via programming questions is recommended [23, 56, 74, 90], it has limitations: (a) overusing such questions could lead to automatically responding *correctly* without paying attention to the questions [90]; (b) in studies such as ours where the software teams include a variety of roles (e.g., product manager, QA, etc.) as well as with specific programming skills (e.g., JavaScript developers), having programming knowledge questions may bias the participants' pools towards more experienced developers in larger companies with traditional programming knowledge, preventing recruiting *novice* developers and those in other SDLC roles; and (c) AI tools like ChatGPT [15] are widely accessible and can handle code-based questions. Thus, these questions are no longer a strong barrier to screen participants. Our analysis of Danilova et al.'s [23] pre-screening questions with GPT-3.5 shows that the tool can answer the questions with 95% accuracy. We discuss how we mitigate these issues below and in Section 4.

Conducting the Survey Prolific maintains a pool of active participants who are regularly screened and vetted by the platform. In our survey, we decided on the sample size based on Prolific's guidelines (a minimum of 300 for a representative sample). We initially pre-screened the Prolific participants based on the following requirements: (a) to be at least 18 years old, (b) fluent in English, and (c) working in industries such as Graphic Design, Information Services, Data Processing, Product Development, Software, Video Games, etc. We used their industry (rather than their role) as a filter, since Prolific does not allow selecting participants based on role. We paid

an average of $25.17/hr to those who completed the survey. After the initial pre-screening, we recruited 686 participants across both US and non-US pools. Out of the 686 participants who started the survey, 14 did not give their consent, and 295 did not finish the survey; hence, they were excluded from our analysis. We then conducted another filtering process to ensure that the participants work in the software industry and, in fact, have software development experience. We asked them, "Q4. In short, tell us about your product and who your customers are." We manually evaluated their responses and cross-checked them with Q6 (their post-secondary degree) and Q9 (their roles). We found that most of them are involved in software development activities such as *"I make a productivity app for Mac & Windows to record & share the user's screen."* We eliminated 15 participants as we could not verify their involvement in SDLC; for example, those with responses as *"NA"* or *"I sell home decor items. My customers are primarily women."* Following these steps, we ended up with a total of 362 participants for our final count.

3.2 Pilot Survey: University Students

Our pilot survey participants were our university's graduate students (who mostly have industry experience through internships and part-/full-time jobs) over the age of 18 from the disciplines of Computing and Information Science, Electrical and Computer Engineering, and Business, who had experience in software development, IT, or related fields. To maintain their anonymity, we did not collect any personally identifiable information such as their contact, names, or company names.

The goal of the pilot study was to gather initial insights and feedback before the deployment of our main study on Prolific. Upon the IRB approval, we launched the survey using Qualtrics. The survey consisted of 40 questions, including 13 short and 27 multiple-choice questions, which were derived based on our informal discussions with developers in small companies and prior gaps in research. We estimated that the survey takes ∼30-40 minutes to complete. Every participant was presented with the same set of questions regardless of their role on a software team. We used the responses to improve our large-scale survey (i.e., Subsection 3.3).

We received 45 responses but most were incomplete due to the survey's length and the diversity of questions. After discussing the study with the participants, we revised and shortened the survey based on participants' role in the SDLC.

3.3 Software Teams Survey

The main feedback we received from the pilot study was that the survey required too much time to complete (∼27 minutes). To address this limitation and to focus on capturing participants' perspectives related to their SDLC roles, we separated the survey questions according to the roles. This shortened the survey duration by 12 minutes and enhanced

the quality of the responses we received. We first asked all participants the same set of 10 questions that are partly related to demographics and the degree of privacy understanding. We then divided the remainder of the questions into four groups, one for each role defined in Table 1. Our breakdown loosely follows the SDLC phases, but we separated the Information Security/Privacy (ISec) roles from the Software Developer (SD) roles to evaluate the significance of security or privacy knowledge in our survey. Although "Others" role was an option, none of the participants selected it.

3.4 Survey Questions

The survey includes a mix of demographic, perception, experiential, and behavioral questions which are crafted based on our RQs (see Section 1) and the challenges identified in prior research regarding creating privacy-preserving applications, such as understanding privacy concepts [2, 9, 40], knowledge of regulations and establishing compliance [4, 28, 33], creating consistent and accurate privacy policies [12, 34, 52, 60, 69, 71, 72, 77, 96, 99], knowledge of privacy approaches and existing tools [16, 38–40]. The complete list of questions (except questions 1-3, which are the required Prolific identification questions and our consent form) is found here.[1]

Demographic questions collect basic information about the participants, such as age, education, their SDLC role, and the company size; e.g., "What areas/roles of the development team are you currently involved with?".

Perception questions aim to understand participants' perceptions toward privacy; e.g., "How do you define privacy?".

Experiential questions ask about their experience with privacy challenges and tasks; e.g., "What was the process for the Privacy Impact Assessment, and who was involved?".

Behavioral questions ask about the participants' behaviors and knowledge related to privacy; e.g., "List any privacy-by-design strategies you have used or know."

4 Ethics & Limitations

Ethical Considerations This research adheres to our university's ethical guidelines and was conducted with our Institutional Review Board (IRB)'s approval. All participants agreed to a thorough consent form that included information about the investigators, the risks, benefits, compensation, and confidentiality. All participants were informed about their voluntary participation, maintaining their right to withdraw at any time. No personally identifiable information was collected, and measures were in place to ensure the anonymity, confidentiality, and security of responses. The contact information of all investigators and the IRB team was also included. No participants contacted the investigators or the IRB about the study or the compensation.

---

[1]Survey questions: http://tinyurl.com/2p9n49e4

Limitations Like most survey studies, our analysis is based on participant self-report data and is affected by self-report bias, recall bias, and social desirability bias. Participants were informed during consent that the survey pertained to privacy due to our institutions' IRB requirement. This may introduce priming and self-selection biases. There is also recruitment bias as the Prolific user base may not fully represent the diverse population of SDLC individuals. We used multiple screening questions to ensure that recruited participants have experience in software development activities (Section 3.1). We adopted a conservative process to remove participants for whom we could not verify their SDLC involvement, however, we may have removed a few professionals. We also asked follow-up and write-in questions to ensure the multiple-choice questions were backed up with written facts. To mitigate the potential for survey responses being generated by AI tools like ChatGPT [15], we minimized open-ended questions in favor of multiple-choice formats and carefully scrutinized the write-in responses to remove those that appeared AI-generated. Short responses with typos and errors suggested that our responses were not AI-generated. Despite our efforts, AI-generated responses could affect the study's outcomes.

We carefully framed our questions so as not to prompt biased responses. However, we could not avoid one leading question that asks about the confidence in their companies' privacy and security measures. We aimed to reduce the bias by providing four options instead of a 'yes' and 'no', with the option to not answer. Additionally, the question follows their own definition of privacy, further helping minimize bias. We employ statistical analyses (like the chi-square test [37]) to ensure the broad applicability of our findings. To control for Type I errors in the presence of multiple hypothesis tests, we report our results after employing Bonferroni correction.

5 Study Analysis Process

Our survey results are organized around our research questions (RQs, see Section 1), focusing on various areas of privacy within the SDLC and across different roles. Our RQs examine the perceptions held, privacy experience and challenges, and privacy behaviors while considering the demographic breakdown (see Section 3.4) to provide additional context and to allow for a more nuanced understanding of the data. Our analysis follows a mixed-method approach, encompassing both quantitative and qualitative methodologies.

Qualitative Analysis We evaluate the descriptive and open-ended questions through open coding procedures and iterative processes. However, in our analysis, we used taxonomies and categories based on the current literature to classify the responses. For the open-ended question regarding the *definition of privacy*, the first two authors, independently, classified 50 responses based on the taxonomy of privacy introduced by Solove [79] and the examples and hypotheses from [41, 50]. Similarly, for the *usage of PETs*, we used PETs categories

from the literature [22, 62, 76]. The first two authors independently assigned categories for the first 25 responses. They then discussed their results, resolved the discrepancies, and created a guideline (see Appendix K). They continued with the rest of the responses, evaluated the agreements and resolved the disagreements in another round of discussion. Lastly, a third privacy expert examined the results to ensure their correctness and completeness. For the non-subjective descriptive questions e.g., *which PbD strategies they use*, one author categorized them based on the current literature, (e.g., privacy by design strategies [47], phases and roles in the SDLC [73] for PIAs) and the second author reviewed them for correctness.

Quantitative Analysis For the questions where our goal is to understand if a correlation exists between the demographics and the privacy-related perception, experience, and knowledge, we conducted statistical analyses. We used the Chi-Squared test [37] to determine whether there is a significant correlation between two categorical variables. For questions where the responses are on a Likert scale, we used the Kruskal-Wallace test [14]. For *perception*, *experience*, and *behavioral* questions, we hypothesize from our RQs that the size of the company, the presence of a CPO or a similar role, the education level, roles, and participants' location may impact their confidence in privacy/security measures, various privacy practices (such as the creation of PIA or privacy policies), and their familiarity with PETs, regulations, and usage of forums. To control Type I errors and avoid false positives, we use Bonferroni correction [75]. Since Bonferroni correction is very conservative and may increase Type II errors, we discuss the results with respect to $alpha = 0.05$ as well as the adjusted value (i.e., $\frac{0.05}{24} = 0.0021$, for our 24 statistical tests).

6 Findings

6.1 Survey Demographics

In our main study, we received a total of 362 responses (after filtering - see Section 3.1). 189 participants reside in the US and the other 173 come from 22 other countries (see Section 1). Table 1 shows a breakdown of participants' roles, with the majority (~55%) in SD roles. As shown in Appendix B - Table 10, most participants identify as male, are below the age of 45, and have completed their BSc., with ~61% in Computer Science (CS), Information Technology (IT), Data Science (DS), and Electrical & Computer Engineering (ECE) majors. This value includes the answers to "Others, please specify". Among those with a Business degree, 61% are in AD (e.g., product manager), and 28% are in SD roles. Among those in the "Other" degree category, 48% identified as SD, 19.5% as QA, 21.0% as AD, and 11.5% as ISec. The company sizes of <100 and 100+ employees are distributed almost equally.

6.2 Perceptions of Privacy

We seek to understand software teams' privacy comprehension by examining how they define privacy, their confidence in their company's practices, and if these differ based on roles or organization differences (i.e., RQ1&2).

6.2.1 Definition of Privacy

One of our key questions is, "How do you define privacy?". The responses were diverse, showing differing perceptions. Some participants defined privacy in terms of data security, highlighting the need to protect user data from unauthorized access. For example, one participant explained that *"It involves implementing measures to safeguard sensitive information, such as encryption, access controls, and data anonymization"*. Others described privacy from a user rights perspective: *"I define privacy as the ability to control all that is related to my information and to keep it from reaching someone who is unauthorized"*. Few responses incorporated legal compliance, with one participant defining privacy as: *"This involves being compliant with regulations and ensuring all data is protected with a least-privilege access model with ownership of the different part data sources with assigned data stewards"*.

To categorize the diverse definitions of privacy, we utilized Solove's taxonomy [79], that breaks down privacy into various categories based on the types of harm of a privacy breach. We chose Solove's taxonomy for two key reasons: (a) it provides a structured and detailed approach to understanding and analyzing definitions of privacy, which is essential with our wide range of definitions and perspectives; (b) it has been widely recognized and used in privacy research [8, 10, 46, 100]. We followed an open coding procedure to map the provided definitions with the taxonomy, as described in Section 5. Multiple classes for each definition were also possible. Figure 1 shows the mapping. (For a breakdown of Solove's Taxonomy see Appendix C - Table 11; the 'Blackmail' category did not apply to any participant's definitions.)

Figure 1 shows that the top frequently occurring categories are 'Disclosure', 'Increased Accessibility', and 'Insecurity'.
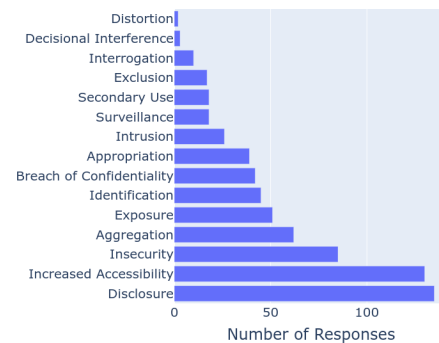
Figure 1: Privacy Definitions based on Solove's Taxonomy

Table 2: Distribution of Participants' Confidence

Role	Yes	No	Unsure	PnS
AD	53 (75.7%)	6 (8.6%)	11 (15.7%)	0
SD	149 (75.3%)	8 (4%)	35 (17.7%)	6 (3%)
QA	25 (63%)	2 (5%)	12 (30%)	1 (2%)
ISec	44 (81.5%)	2 (3.7%)	8 (14.8%)	0
Total	**271**	**18**	**66**	**7**

This result indicates that most participants either consider the traditional definition of privacy as *control over personal information* or perceive privacy in terms of *security*. For 'Disclosure', one participant highlights the importance of transparency and clear communication about data collection purposes and user control: *"Privacy is the assurance that all data belonging to an individual will be disclosed to others only with that individual's consent, for uses understood and approved by that individual."* For 'Increased Accessibility', a participant who works with genetic data underscores the need for controlled access to such information, only granting access if needed: *"The users' ability to define who can access their data and even in that what kind of data can be accessed. As I work in genetic data from patients in my line of work, the clinical information is always controlled access and only researchers working on the particular project can gain access on a need-to-know basis."*

We further examined how privacy perceptions differ across various roles. Almost 50% of participants in AD or ISec roles define privacy as 'Disclosure', while QA and SD roles mostly consider privacy as 'Increased Accessibility', which is related to *access control*. ISec roles mentioned 'Aggregation' more frequently than other roles, which is an anonymization technique used only in privacy rather than security.

> The variety in our participants' definitions of privacy shows the complexity of privacy perceptions, and the need for a holistic approach that covers a variety of aspects of privacy throughout the SDLC.

6.2.2 Confidence in Security and Privacy Measures

We asked participants about their confidence in the privacy and security measures implemented in their organization. Table 2 shows the distribution of the participants and their responses. Note that 'PnS' stands for 'prefer not to say'. In all roles, most participants are confident in their company's security and privacy measures. Interestingly, ISec members are the most confident while the QA members are the most uncertain. This can be either due to QA members considering privacy and security as an afterthought [40], thus ignoring these requirements, or because they encounter more non-compliance instances during testing than any other roles.

We analyzed whether there is a correlation between participants' confidence in security and privacy measures and

Table 3: Distribution of Company Size vs Existence of a CPO

Company Size	Yes	No	Unsure	Others
0–20	31.5%	51.5%	15.7%	1.4%
21–100	46.1%	29.2%	21.6%	3.1%
100+	47.3%	34.9%	17.8%	0%

their demographic factors, such as the company's size (**H1a**), participants' roles (**H1b**), education level (**H1c**), and the presence of CPO or a similar position (**H1d**) (see Appendix D and Table 12 for more details). As shown in Table 12, with Bonferroni adjustment ($\frac{0.05}{24} = 0.0021$), we cannot reject the null hypothesis for **H1a**, **H1b**, and **H1c** ($p-value = 0.494, 0.654$ and 0.570); thus, we find no correlation between confidence in security and privacy measures and a company's size, participants' roles, or education levels. However, with a $p-value = 0.0007$ for **H1d**, we can reject the null hypothesis and say there is a correlation between the presence of a CPO (or similar position) and confidence in privacy and security measures. We further evaluate whether the existence of a CPO could lead to positive privacy outcomes in Subsection 6.2.3.

We asked the ISec members specific questions regarding their company's security and privacy measures. When asked *"whether their company conducts security audits for third-party software used in their products"*, slightly more than half (~56%) said 'Yes' while a large number (~38%) were 'Unsure'. This is alarming since research shows a large number of third-party software and libraries include security and privacy vulnerabilities [1, 42, 97]. However, when we asked *"whether their company securely manages encryption keys and implements encryption algorithms and access control policies"*, more than 70% responded 'Yes' – which highlights inconsistencies in privacy practices even among experts.

> A CPO is important in fostering employees' confidence in the privacy and security measures of an organization.

6.2.3 Presence of a Chief Privacy Officer (CPO)

To evaluate the impact of a CPO or other similar roles on privacy practices, we focus on the AD and SD roles, who are the majority of our participants (i.e., 268 (74%)). We did not include ISec and QA teams to avoid any response bias, due to their active privacy role in the company. We asked them "Do you have a Privacy Officer or similar position in your company?". Table 13 in Appendix E shows the distribution. Interestingly, only slightly more participants responded 'Yes' (42.6%) than 'No' (38.4%). ~18% were 'Unsure', which may indicate the lack of proper communication among employees regarding the company's privacy practices and the purpose of a CPO. 1.1% responded 'Other', which included a legal team or a CTO. Among those that said 'Unsure', 23% are in AD and 77% are in SD roles which may indicate CPO members communicate more with the management team (i.e., AD).

We investigated whether the larger companies have a CPO. Table 3 shows the distribution of the presence of a CPO based on the company size. Here, we see the presence of a CPO increase with the company size. We also observe that companies of all sizes have a sizable number of 'Unsure' responses.

We further asked the participants "When you have a question about compliance with regulations, what do you do?". The participants could select more than one option. Appendix E - Table 15 shows the distribution of the responses. About half of the respondents (50.1%) mention they ask lawyers or a CPO, while 23.1% look at the best practices and standards (such as NIST guidelines), and 18.5% use developers' forums (such as Stack Overflow). Among 'Other' sources, they mainly mention 'search Internet' or 'ask a colleague'.

We analyzed whether the existence of a CPO (i.e., access to a legal or privacy expert) could impact the creation of PIA (**H2a**), the familiarity with PETs (**H2b**), the number of privacy breaches (**H2c**), or is influenced by the company size (**H2d**). Appendix E and Table 14 show the list of the hypotheses and the results of the tests. With Bonferroni correction, our results show that the presence of a CPO correlates with the size of a company ($p-value < 0.00001$). This correlation indicates that larger companies are more likely to have a CPO or a legal/privacy expert to help mitigate privacy risks, which is aligned with findings in [7]. However, with the p-value adjustment, we do not find a correlation between the presence of a CPO (or a similar position) and the creation of a PIA ($p-value = 0.1005$) and the use of PETs ($p-value = 0.008$). This may not be surprising, especially since a majority of SD roles, who are the main users of PETs and are involved in the PIA creation, are unaware of a CPO role. Our analysis also did not reveal a significant correlation between the presence of a CPO and the number of privacy breaches experienced by the organization ($p-value = 0.359$). This suggests that the presence of a CPO, while important and necessary, may not be sufficient to help minimize privacy breaches.

> Although a CPO could improve confidence in a company's privacy measures, it has limited effectiveness in enhancing privacy practices and reducing breaches.

6.3 Experience with Privacy

We ask members of software teams in various roles about their *experience* with creating privacy policies and/or PIA, as well as practices to ensure the protection of users' data to better understand their privacy challenges (i.e., RQ2&3).

6.3.1 Creation of a Privacy Impact Assessment

A Privacy Impact Assessment (PIA) is a critical tool for identifying and mitigating privacy risks at any stage of software development. Recently, PIAs and their variations, such as the Data Protection Impact Assessments (DPIAs), have become
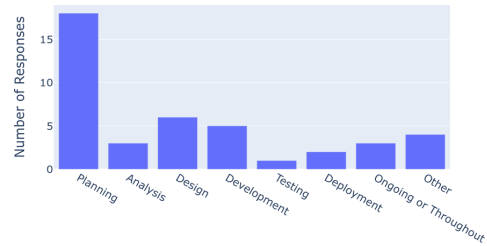
Figure 2: Stages of the SDLC When PIAs are Created.

a requirement in GDPR [27] and CCPA [35]. This tool allows organizations to address privacy and security issues before they become problems. In our survey, we asked participants if a PIA was created at any time throughout development, and if the answer is yes: at what stage it was created, who was involved, and what the process for creation was.

We received 311 responses, where only 43 (14%) of them (where more than half were outside of EU+UK) reported that they created a PIA at any point in the SDLC, while a significant proportion (57.2%) reported that they did not (see Appendix F - Figure 5). This indicates a lack of awareness regarding the existence or the need for PIAs (i.e., the PIAs are non-existent or are conducted without their knowledge by the CPO or other teams). We also observed that ∼25% are unsure about whether a PIA was created, which may highlight a gap in communication within a company about its privacy practices. ∼4% chose 'Prefer not to say'.

Among the 43 who created a PIA, 3 (∼7%) did not answer the follow-up questions. Our results show that PIAs were created at various stages in the SDLC (see Figure 2), but ∼51.0% are at the planning and analysis stages. One participant who reported that a PIA was created during the planning stage said, *"At the start of development of idea because privacy is more important than all things"*. Some participants reported creating a PIA at the start of development, e.g., *"We created a [PIA] at the beginning of the software development process. This allowed us to identify potential privacy risks and develop strategies to mitigate them"*. Others mentioned during the design, or even towards the end of development.

The sizable number of participants (42%) involved in PIA during the later stages in SDLC may indicate that privacy requirements are not considered early on, and are only included as an afterthought – which is aligned with the findings in [40]. Furthermore, the variation in the timing of the PIA creation shows the need for a more standardized approach to incorporating privacy considerations into software development.

Figure 3 shows the distribution of the roles involved in PIA creation. More than one category was allowed. The responses are also diverse. Some participants reported that they created the PIA themselves or it was a team effort (i.e, SD & QA teams), while others reported that it was done by the CPO or external Legal team, ISec teams, upper management (i.e., CEO or CTO), or even the client (External). This shows that
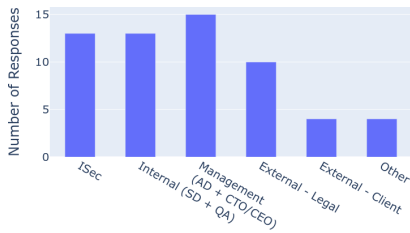
Figure 3: Distribution of Roles Involved in PIA Creation.

the responsibility for privacy can be distributed across various roles, which again highlights the need for clear communication, collaboration, and defined privacy practice processes.

The responses to the question regarding the process for creating the PIA also varied. Some initiate the process by downloading a template and collaborating with internal specialists, while others seek guidance from lawyers, executives, or third-party experts. A common approach involved consulting with professionals, with one participant mentioning that they *"outsourced [a] developer that specialises in data privacy and security"*. Several participants mentioned the involvement of specific roles, such as the Data Chief, IT teams, and privacy protection specialists. The process often involved cross-functional teams. In some cases, senior leadership, e.g., the CTO, CEO, or owner, played a pivotal role in the process.

Lastly, we evaluated the PIA correlation between the company size (**H3a**) and the participants' confidence in privacy and security measures (**H3b**) (see Appendix F). The results show a significant correlation $p-value < 0.0001$ for both tests - even after Bonferroni correction.

> Most members of software teams are not familiar with PIA or are unaware of its creation. However, those involved in PIA emphasize the need for its creation in the initial phases of the SDLC in a collaborative process with experts from various departments and consultants.

6.3.2 Creation of Privacy Policies

Privacy policies describe how, why, and how long an application uses personal information. Regulations [27, 35] and the FTC [92] require companies to provide users with detailed privacy policies. Research shows that these policies may be inconsistent with apps [78, 99], since they are either created by outside legal experts (who may not fully comprehend the apps) or by using privacy policy generators [98].

We asked the AD team about their experience and challenges with privacy policies (as other roles are often only indirectly involved). Out of the 70 participants, 3 did not provide any answer. Of the rest, only 11 (17%) have been involved in the creation of a privacy policy, and they used 'legal experts' the most (64.0%), followed by 'templates' (45.5%), and 'privacy policy generators' (36.4%). More than one response could be selected. Two of them mentioned that they

either 'search the Internet' or 'ask for team input', in addition to using privacy policy generators and templates. In 60% (out of 45.5%) of cases that used 'templates', and in 50% (out of 36.4%) of cases that used 'privacy policy generators', a 'legal expert' has also been selected. This result matches with prior research that legal experts in a company are mainly involved in the privacy policy creation, which may lead to inconsistencies between the app and the policy [78]. We also noticed that those who said 'Yes' are mostly from companies with less than 100 employees (∼64%) and with a CPO (∼55%).

Finally, we asked the 11 participants who responded 'Yes', "What challenges did you face when creating your privacy policy?". We received 10 responses. Six of them describe the challenges regarding compliance with regulations in multiple international jurisdictions, and understanding legal jargon, rules, and standards. One specifically had concerns regarding compliance, since they use privacy policy generators: *"...differences between different countries and their requirements since we are international."* Five respondents describe their main challenge as ensuring completeness (i.e., covering all personal information), soundness, and language of privacy policies. E.g.: *"Whether the wording I chose was going to cover all the bases I needed it to and whether it was clear and easy to understand."* or *"Which rules and text to inform users;..."* One of those five respondents was also concerned about which template to choose. Four others did not find the process challenging since they trusted the legal expert to help.

> Compliance with regulations, and ensuring completeness and correctness are among the most common challenges in creating a privacy policy. Software teams use several tools besides legal experts to help create privacy policies.

6.3.3 Privacy Practices to Protect Users' Data

We tailored some of the privacy practice questions based on the role, specifically for ISec, SD, and QA teams. We asked ISec members: "How do you ensure that data collected from users is used only for intended purposes?". They discussed various approaches. ∼32% emphasized the importance of documentation to ensure transparency and accountability, with one noting *"the meticulous documentation of every step in the data usage process"*. Encryption emerged as a common theme, with participants mentioning sending encrypted documents and ensuring data is stored securely. A respondent states *"I would send documents encrypted and compressed into a zip file, and instruct them to delete the file once the information is accessed."* Limiting access to data is another frequent approach, with 30.2% stressing the importance of restricting data access to only those who need it and maintaining logs to track any access. 16.98% stated the significance of transparency, ensuring they only collect necessary data, obtaining user consent, and regularly monitoring data usage. A few (9.43%) pointed out the importance of adhering to spe-

cific regulations, such as the Health Insurance Portability and Accountability Act (HIPAA) [43] and the Family Educational Rights and Privacy Act (FERPA) [64]. 16.98% admitted to not having direct control over data but trusted their organization's protocols and training to handle data responsibly.

We also asked the same group "How do you manage access to sensitive user data in your organization?". Role-based access controls, multi-factor authentication, and encryption are common strategies employed to safeguard sensitive information. One respondent shared, "*We limit access to systems based on who really needs to access that data.*". Such measures ensure that only authorized personnel can access sensitive data, thereby minimizing potential breaches. Regarding data retention practices, only 47.17% of respondents state that they have been involved in removing user data either after its predetermined lifespan or upon user request.

We asked the SD members: "If you encounter a privacy concern at any point in the software development process, what steps would you take?". More than 95% of them take the concerns very seriously. For example, one participant mentions *"run a risk assessment"* and another mentions *"We take the app offline and start iteratively testing parts of the app to see where the privacy concern is."* About 36% deal with the concern internally to fix it and communicate it with the client and upper management. Another 35% directly escalate it to their supervisors, while 20% seek help from the ISec team or lawyers. A handful contact the client first.

We asked the QA team: "How do you verify that third-party systems used in your products are privacy compliant?". Similarly, we received diverse responses. Only 56.6% confirmed that their companies conduct security audits of these third-party systems. Some mentioned the significance of conducting vulnerability assessments and penetration testing to ensure third-party systems' compliance (23%). Some respondents discussed that they rely on reading privacy policies and contracts of third-party systems (28%), while others emphasized the importance of legal agreements and monitoring data transfers (15%). About 22% admitted to not being directly involved in this process, placing trust in their organization's legal and security teams, which is aligned with findings in [19].

Lastly, regarding the QA teams' practices for testing for privacy breaches and data leaks, they emphasized the importance of understanding the data they work with and always being vigilant about potential breaches. Regular manual or automated testing is a common theme. ~27% of them mentioned the use of penetration testing, both internally and via third-party services. Others stressed the importance of using fake data during testing phases and ensuring that real user data is always encrypted and protected. ~16% of respondents admitted to not being directly involved but trusted their organization's protocols and cybersecurity measures.

> The most common privacy practices among SD, ISec,
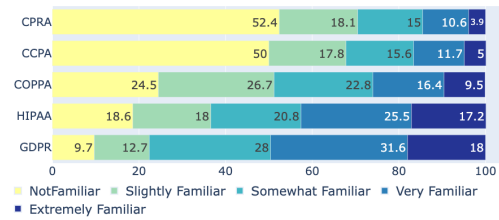
Figure 4: Familiarity with Different Regulations.

> or QA teams are documentation, auditing, and security techniques (such as access control and encryption). QA teams rely heavily on legal and ISec teams to ensure data protection and are less involved themselves.

6.4 Privacy Awareness and Behaviors

We assess privacy *behaviors* based on familiarity with regulations, PbD, PETs, and such knowledge sources (i.e., RQ4).

6.4.1 Familiarity with Privacy Regulations

In recent years, several regulations have been introduced that developers need to comply with. Non-compliance with these regulations may lead to financial penalties, sometimes up to 4% of the annual turnover of the company [27]. However, these regulations include legal terminologies that may not be familiar to members of the software teams. To understand the degree of familiarity and awareness, we asked all 362 participants about their familiarity with GDPR, HIPAA, COPPA, CCPA, and the California Privacy Rights Act (CPRA). The answers are on a Likert Scale (see Figure 4).

We combined the results from 'somewhat familiar', 'very familiar', and 'extremely familiar' together and found that software teams' members, regardless of their region and roles, are more familiar with GDPR (77.35%) and HIPAA (63.26%). COPPA, CCPA, and CPRA are all 50% or below. ISec teams are the most familiar with all regulations among all roles, followed by the SD and AD teams. The QA teams are the least familiar with 7.5% familiarity with CCPA and CPRA, and 65.0%, and 57.5% with GDPR and HIPAA.

We asked participants "How did you learn about the previous regulations?". More than one option could be selected. As shown in Table 4, the majority are self-taught while university education ranks second. Among all roles, the ISec team has the highest percentage of learning about regulations through university education (33.3%), which is more likely through cybersecurity courses. We also asked the participants to describe the other sources they used to learn about privacy regulations. In most cases, they mentioned 'training at work' as the source; however, 2 participants mentioned 'social media' and 'YouTube' as their source.

Table 4: Distribution of Participants' Learning Experience

Role	Self Taught	Lawyer	University Education	IAPP Cert.	Others
AD	57.1%	5.7%	20.0%	1.4%	15.8%
SD	56.6%	3.5%	21.7%	2.5%	15.7%
QA	85.0%	7.5%	2.5%	0.0%	5.0%
ISec	42.6%	3.7%	33.3%	9.3%	11.1%
Total	**57.7%**	**4.4%**	**21.0%**	**3.1%**	**13.8%**

> GDPR is the most familiar regulation among all participants due to its comprehensiveness. ISec teams are more likely to learn about regulations through university education; hence, are more familiar with them than other groups. QA teams are the least familiar.

6.4.2 Familiarity with Privacy by Design (PbD)

Privacy by design (PbD) strategies introduced by Hoepman et al. [45] have gained interest in helping developers to be compliant with regulations. We asked the SD members (i.e., 198 participants) if they are aware of PbD, and if they answered yes, whether they used them (see Appendix G - Table 17) and to list the ones they used. ∼46% are familiar with PbD approaches while ∼25% are unsure, which indicates the potential knowledge gap and opportunity for educating developers. Out of those who answered 'Yes' to the awareness of the PbD question, only 57.1% had employed such strategies in their work. Of the remaining, 23.1% did not use them and 16.5% were unsure. This result suggests that even among developers who are familiar with such strategies, not everyone acts on this awareness – which may indicate the lack of usability and readiness of PbD for day-to-day developers' tasks [88] or other organizational factors, such as lack of resources [51].

Lastly, we evaluated the responses about the usage of specific PbD strategies (multiple answers were possible). Interestingly, our results are aligned with the findings of Tahaei et al. [87] (see Table 5). Our top categories are 'hide' (22), 'minimize' (21), 'inform' (17), and 'control' (12), while 'enforce' (1) and 'abstract' (2) are rarely discussed. One participant mentions *"Mostly minimise. Its the most straightforward."* This response reinforces our result in that 'minimize' is one of the easiest strategies to implement. We also received responses regarding Anne Cavoukian's PbD principles [17] such as 'privacy by default' (6 times) and 'proactive' (twice). The use of PIA was also mentioned 5 times as a strategy.

> Our findings show that PbD approaches are not yet commonly used, and their lack of adoption underscores the gap in developers' knowledge regarding PbD and their usability in day-to-day developers' tasks.

Table 6: Usage of PETs in Software Development Process

Privacy Enhancing Technology (PET)	Percentage
Encryption	70.48%
Access Control/Identity Protection	34.29%
Anonymity and Pseudonymity	9.52%
Differential Privacy Approaches	8.57%
Secure Communication/VPN	8.57%
Privacy-Enhanced Anti Web Tracking	0.0%

6.4.3 Use of Privacy-Enhancing Technologies (PETs)

Using PETs is another critical component of privacy protection, that allows better protection and maintenance of data privacy against outside threats. We asked the SD team, who are the main users of PETs, if they used any PETs, and if so to list them. Out of the 198 participants, 2 did not respond. 111 of them (56.63%) mentioned that they use some PETs while 36 (18.37%) do not. About 25% are unsure. These results are almost aligned with the degree of familiarity and usage of PbD. There was an increase (∼10%) in PETs familiarity and/or usage in comparison to PbD, which shows that these technologies are more common and tangible for developers, especially those related to encryption and access control. We grouped responses into 6 categories shown in Table 6 (definitions in Appendix K). Encryption and access control, which are primarily security-focused, were the most common, followed by anonymization methods and differential privacy.

Lastly, we investigated the correlations between the PETs' familiarity and the company size (**H4a**), confidence in security and privacy measures (**H4b**), and education level (**H4c**) (see Appendix H). With the adjusted p-value, we find no correlations ($p-value = 0.254$, 0.529, and 0.704, respectively).

> PETs are slightly more commonly used than PbD strategies. However, there is still a gap in their familiarity, where more than 40% of developers do not use them or are unsure of their usage. The most commonly used PETs are more security-oriented concepts, than privacy.

6.4.4 Developers' Sources for Privacy Information

As discussed in Section 2, developers sometimes seek privacy-related guidance on forums, such as Reddit or Stack Overflow (SO). We asked the SD teams how often they use various developers' forums for their privacy-related questions. Table 7 shows the distribution of the responses and their frequencies. ∼70% and ∼58% of the respondents use SO and GitHub *at least 1-3 times per month*, while for Reddit and Quora, this number is about 34.5% and 18.5%. About 57% of the respondents find these forums very or extremely useful, while less than 6% find them not useful at all. In cases where they do not find the answer on these forums, the SD team discusses their questions with the security or privacy experts, asks their

Table 5: Distribution of PbD Strategies Used by Developers

Minimize	Hide	Separate	Abstract	Inform	Control	Enforce	Demonstrate
21	22	7	2	17	12	1	4

Table 7: Frequency of Usage of the Developers' Forums

Forums	Never	Rarely	1-3/M	1-3/W	Daily
SO	13.1%	17.1%	26.1%	24.1%	19.6%
GitHub	18.4%	23.9%	23.4%	19.9%	14.4%
Reddit	30.5%	35.0%	20.0%	10.5%	4.0%
Quora	54.5%	27.0%	12.5%	5.5%	0.5%

Table 9: Distribution of Regulations Familiarity

Location	GDPR	HIPAA	COPPA	CCPA	CPRA
US+CA	71%	84%	53%	48%	44%
EU+UK	89%	37%	38%	11%	9%
Others	69%	51%	57%	29%	29%

teammates, or uses AI tools. In Appendix I, we provide a more detailed analysis regarding the usage of the forums.

> Developers often seek privacy-related information from online forums, where more than 50% of participants use either Stack Overflow or GitHub at least 1-3 times per month and they find these forums useful.

7 Location Analysis

Our large-scale survey has responses from the US (189 responses) and non-US (173 responses from 22 countries: EU+UK, South Africa, Canada (CA), Mexico, and Chile), enabling us to examine differences in perceptions, experiences, and behaviors. We group the countries into three regions based on their similarities in privacy regulations: US+CA (192), EU+UK (132), and 'Other' countries (38). To evaluate the difference in *perception*, we examine whether participants' location correlates with their confidence in privacy and security measures (**H6a** in Appendix J) and the presence of a CPO (**H6b**). Both hypotheses do not hold ($p-values$ are 0.0567 and 0.6470). Table 8 shows the presence of a CPO across the three regions. The percentage of 'Yes' is almost equal between US+CA, EU+UK, and the 'Other' countries, while slightly more US+CA participants mentioned "no CPO" than elsewhere. This is not surprising since GDPR, the UK Data Protection Act of 2018, and the US HIPAA (Art.164.530) all require having a privacy officer or officials in a similar role.

We evaluated whether there is a significant difference between participants' *experience* in the three regions regarding the creation of PIA (**H6c**) and the number of privacy breaches (**H6d**). With $p-value$ 0.7724, we find no correlation for PIA.

Table 8: Distribution of Location-based CPO Presence

Locations	Yes	No	Unsure	Others
US+CA	43.7%	41.5%	14.1%	0.7%
EU+UK	41.7%	36.1%	20.3%	1.9%
Other Countries	43.5%	30.4%	26.1%	0%

However, there is a correlation between the regions and privacy breaches ($p-value = 0.0010$). We also analyzed the *privacy behaviors* in the three regions concerning familiarity with PbD (**H6e**) and usage of PETs (**H6f**). With $p-values$ 0.3120 and 0.8588, we do not find any correlation that suggests that usage of PETs and PbD are equally (un)common in all regions. Since participants are from regions governed by different privacy laws, we investigated their familiarity with CCPA [35] (**H6g**) and GDPR [27] (**H6h**). As expected, we find a significant correlation between the participant's familiarity with the two regulations ($p-values$ are < 0.0001 and 0.0009 respectively). Due to the global reach of many apps, SDLC teams are responsible for complying with various regulations. We further evaluated the responses to the familiarity with each regulation in various regions. We combined the responses given for at least *somewhat familiarity* (i.e., somewhat, very, extremely familiar) and found that participants in the US+CA are most familiar with HIPAA while the rest are most familiar with GDPR. Those from 'Other' countries are also more familiar with the US regulations than those residing in the EU+UK. Table 9 shows the distribution.

8 Discussion

Summary of Findings Concerning *privacy perception*, our survey identifies that the majority of the participants define privacy in terms of control over personal information and disclose only when needed, or in terms of security. In other research [48, 85], data protection and security were the most common definitions. Having a CPO or a similar role positively impacts confidence in protecting users' data. However, we found out that a sizable portion of the participants are unaware of such a role in their company, which may lead to ineffectiveness in utilizing privacy tools or reducing privacy breaches. Lack of proper communication among various roles is a challenge that other research also identified [48, 85]. Our findings also align with [7] and [48], which observed a correlation between company size and having a CPO. However, we did not observe significant location-based differences in these perceptions. This is interesting but not surprising, since GDPR, HIPAA, the UK Data Protection Act, and Protection of

Personal Information Act (POPIA) all require a CPO or similar roles. Several of our US participants mentioned (in Q27) that they collect Protected Health Information (PHI), which falls under HIPAA; e.g., one participant says *"Health related data about people involved with our insurance companies"*. The extensive privacy requirements from these regulations likely explain why we observed no significant geographical differences in terms of participants' confidence, familiarity with PbD, and the usage of PETS.

In terms of *privacy experience*, most participants rely on legal experts to help create privacy policies; unlike [7] where creating a privacy policy was not the priority. Our study also shows that participants are primarily concerned about multi-jurisdictional compliance. Most of them are not involved in creating a PIA. The majority of those involved believe a PIA should be created during the planning or analysis phases; this is almost similar to findings in [40, 48]. Our participants emphasized the importance of detailed documentation regarding the data lifecycle, as well as using encryption and access control tools to protect the confidentiality and integrity of data. Interestingly, the QA teams rely more than others on security, privacy, and legal experts to implement and enforce privacy and security rules. Other studies did not examine the privacy practices of QA roles, separately.

Regarding *privacy behavior*, we identified that less than half of the participants are aware of PbD and an even smaller number use them. Similar to [87], 'hide', 'minimize', 'inform', and 'control' are more commonly used. The usage of PETs is slightly more prevalent than PbD, but the focus is more on security practices, such as encryption and access control; similar to other research that found security concepts are more tangible [7, 40, 48, 85]. Anonymization techniques are not used frequently enough. We also find that although $\sim 53\%$ of our participants are from the US+CA, most are more familiar with GDPR than US-based regulations such as COPPA and CCPA. ISec experts are among the most knowledgeable about various regulations, while QA teams are the least familiar. Other works focus mainly on GDPR and CCPA and do not explore details regarding participants' familiarity [7, 40, 48, 85]. Most participants tend to seek answers to their privacy questions from developers' forums in addition to legal/policy experts; unlike [7] where they used 'friends' or 'social media'.

Research Directions Insights from the related work and our survey results highlight the need for approaches to operationalize PbD strategies and incorporate them into design and development. PbD patterns [93, 94] provide detailed information about their usage and high-level solutions, but still lack implementation. Approaches that detect privacy behaviors in code [53, 55] and further link them to patterns, or leverage automated code generation techniques to generate code from privacy patterns are yet to be explored.

Our survey highlights software teams' challenges in creating accurate PIAs and privacy policies. Research directions that focus on automated approaches to detect the information types, privacy practices, and purposes pre- [49] and post-development [53, 55], or to generate privacy statements from code [54] could alleviate the challenges regarding accuracy, consistency, and compliance.

Developers seek answers to their privacy-related questions from developers' forums, though increasingly use tools such as ChatGPT [15, 67]. However, these tools may not always provide accurate responses [24]. Developing methods to help translate developers' privacy-related questions into accurate privacy code snippets requires further attention [30].

Our survey indicates that software teams face challenges in understanding and adhering to privacy regulations; thus, there is a need for approaches to help better understand such regulations, and establish and maintain compliance. However, most research focuses on detailed requirements analysis, not suitable for agile app development. Future studies could focus not only on automated extraction of legal/privacy requirements but also on generating (privacy-related) user stories to be used in agile development. Research directions on automated approaches to monitor compliance and nudge developers towards compliant approaches are also worth addressing [18].

Educational Takeaway Similar to other work [7, 48, 85], our work shows the need for a more focused educational approach toward privacy in the SDLC. While currently, many courses emphasize security, it is important to tailor specific courses that include advanced privacy topics such as: regulations; the importance of PIA and other artifacts; challenges in privacy policy creation; and approaches such as PbD, differential privacy, and federated learning. This distinction between privacy from security is crucial since privacy encompasses a broad spectrum of concerns, including data handling, user consent, and transparency. Software teams should be equipped with educational modules and tools that foster and support life-long learning of dynamic privacy concepts. Nudging developers towards more privacy-preserving solutions through online support and tools is important. Balebako et al. [6] suggest that with the right guidance, developers can be encouraged to prioritize privacy in their design and development processes.

9 Conclusion

In this paper, we examined privacy perceptions, practices, and behaviors of SDLC team members during software development. Our findings suggest a need for standardized privacy practices, educational awareness and implementation of PbD, and a privacy expert to promote privacy awareness and compliance. We identified gaps in privacy practices among software teams. Finally, we provide research and educational directions to reduce the challenges in implementing these practices.

In the future, we will extend our research to conduct a comparative analysis within the US states. We will also evaluate whether developers over-claim their expertise in a new study. We will look into how privacy is taught at educational institutes, both in computer science and at Law schools.

References

[1] Mahmoud Alfadel, Diego Elias Costa, and Emad Shihab. Empirical analysis of security vulnerabilities in python packages. *Empirical Software Engineering*, 28(3):59, 2023.

[2] Atheer Aljeraisy, Masoud Barati, Omer Rana, and Charith Perera. Privacy laws and privacy by design schemes for the internet of things: A developer's perspective. *ACM Computing Surveys (CSUR)*, 54(5):1–38, 2021.

[3] Noura Alomar and Serge Egelman. Developers say the darnedest things: Privacy compliance processes followed by developers of child-directed apps. *Proc. on Privacy Enhancing Technologies*, 4(2022):24, 2022.

[4] Orlando Amaral, Sallam Abualhaija, Mehrdad Sabetzadeh, and Lionel Briand. A model-based conceptualization of requirements for compliance checking of data processing against gdpr. In *2021 IEEE 29th Int. Requirements Engineering Conf. Workshops (REW)*, pages 16–20, 2021.

[5] Renana Arizon-Peretz, Irit Hadar, Gil Luria, and Sofia Sherman. Understanding developers' privacy and security mindsets via climate theory. *Empirical Software Engineering*, 26:1–43, 2021.

[6] Rebecca Balebako and Lorrie Cranor. Improving app privacy: Nudging app developers to protect user privacy. *IEEE Security & Privacy*, 12(4):55–58, 2014.

[7] Rebecca Balebako, Abigail Marsh, Jialiu Lin, Jason I Hong, and Lorrie Faith Cranor. The privacy and security behaviors of smartphone app developers.(2014). *DOI: http://dx. doi. org/10.1184*, 1, 2014.

[8] Kenneth A Bamberger and Deirdre K Mulligan. Privacy on the books and on the ground. *Stanford Law Review*, pages 247–315, 2011.

[9] Kathrin Bednar, Sarah Spiekermann, and Marc Langheinrich. Engineering privacy by design: Are engineers ready to live up to the challenge? *The Information Society*, 35(3):122–142, 2019.

[10] Colin J Bennett and Charles D Raab. *The governance of privacy: Policy instruments in global perspective*. Routledge, 2017.

[11] J. Bhatia and T.D. et al. Breaux. Privacy risk in cybersecurity data sharing. In *Proc. of the ACM on Workshop on ISCS*, pages 57–64, 2016.

[12] Travis D. Breaux, Hanan Hibshi, and Ashwini Rao. Eddy, a formal language for specifying and analyzing data flow specifications for conflicting privacy requirements. *Requirements Engineering*, 19(3):281–307, 2014.

[13] Travis D. Breaux, Daniel Smullen, and Hanan Hibshi. Detecting repurposing and over-collection in multiparty privacy requirements specifications. In *Requirements Engineering Conference (RE), 2015 IEEE 23rd International*, pages 166–175. IEEE, 2015.

[14] Norman Breslow. A generalized kruskal-wallis test for comparing k samples subject to unequal patterns of censorship. *Biometrika*, 57(3):579–594, 1970.

[15] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[16] Fei Bu, Nengmin Wang, Bin Jiang, and Huigang Liang. "privacy by design" implementation: Information system engineers' perspective. *International Journal of Information Management*, 53:102124, 2020.

[17] Ann Cavoukian. Privacy by design - the 7 foundational principles implementation and mapping of fair information practices. *www.privacybydesign.ca*, 2009.

[18] Checks. Simplify compliance with google. https://checks.google.com/, 2024 (accessed Jun 6, 2024).

[19] Virginie Cobigo, Konrad Czechowski, Hajer Chalghoumi, Amelie Gauthier-Beaupre, Hala Assal, Jeffery Jutai, Karen Kobayashi, Amanda Grenier, and Fatoumata Bah. Protecting the privacy of technology users who have cognitive disabilities: Identifying areas for improvement and targets for change. *Journal of Rehabilitation and Assistive Technologies Engineering*, 7:2055668320950195, 2020.

[20] Michael Colesky, Jaap-Henk Hoepman, and Christiaan Hillen. A critical analysis of privacy design strategies. In *2016 IEEE security and privacy workshops (SPW)*, pages 33–40. IEEE, 2016.

[21] Asmita Dalela, Saverio Giallorenzo, Oksana Kulyk, Jacopo Mauro, and Elda Paja. A mixed-method study on security and privacy practices in danish companies. *arXiv preprint arXiv:2104.04030*, 2021.

[22] George Danezis, Josep Domingo-Ferrer, Marit Hansen, Jaap-Henk Hoepman, Daniel Le Metayer, Rodica Tirtea, and Stefan Schiffner. Privacy and data protection by design-from policy to engineering. *arXiv preprint arXiv:1501.03726*, 2015.

[23] Anastasia Danilova, Alena Naiakshina, Stefan Horstmann, and Matthew Smith. Do you really code? designing and evaluating screening questions for online surveys with programmers. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, pages 537–548. IEEE, 2021.

[24] Zack Delile, Sean Radel, Joe Godinez, Garrett Engstrom, Theo Brucker, Kenzie Young, and Sepideh Ghanavati. Evaluating privacy questions from stack overflow: Can chatgpt compete? In *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)*, pages 239–244. IEEE, 2023.

[25] Cynthia Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer, 2006.

[26] Anirudh Ekambaranathan, Jun Zhao, and Max Van Kleek. "money makes the world go around": Identifying barriers to better privacy in children's apps from developers' perspectives. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.

[27] European Union. The eu general data protection regulation (gdpr). http://www.eugdpr.org/, 2024 (accessed February 10, 2024).

[28] Saad Ezzini, Sallam Abualhaija, Chetan Arora, Mehrdad Sabetzadeh, and Lionel C. Briand. Using domain-specific corpora for improved handling of ambiguity in requirements. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, pages 1485–1497, 2021.

[29] Federal Trade Commission. Children's online privacy protection rule; final rule. http://tinyurl.com/5fh55th2, 2024 (accessed Feb 12, 2024).

[30] Zhangyin Feng, Daya Guo, Duyu Tang, Nan Duan, Xiaocheng Feng, Ming Gong, Linjun Shou, Bing Qin, Ting Liu, Daxin Jiang, and Ming Zhou. CodeBERT: A pre-trained model for programming and natural languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1536–1547. ACL, 2020.

[31] Sepideh Ghanavati, Daniel Amyot, and Liam Peyton. Towards a Framework for Tracking Legal Compliance in Healthcare. In John Krogstie, Andreas Opdahl, and Guttorm Sindre, editors, *Advanced Information Systems Engineering*, pages 218–232. Springer, 2007.

[32] Sepideh Ghanavati, Daniel Amyot, and Liam Peyton. Compliance analysis based on a goal-oriented requirement language evaluation methodology. In *2009 17th IEEE International Requirements Engineering Conference*, pages 133–142. IEEE, 2009.

[33] Sepideh Ghanavati, Daniel Amyot, and André Rifaut. Legal Goal-oriented Requirement Language (Legal GRL) for Modeling Regulations. In *Proc. of the 6th International Workshop on Modeling in Software Engineering*, pages 1–6, New York, NY, USA, 2014. ACM.

[34] Alessandra Gorla, Ilaria Tavecchia, Florian Gross, and Andreas Zeller. Checking app behavior against app descriptions. In *Proc. of the 36th Int. Conference on Software Engineering*, pages 1025–1035, 2014.

[35] Government of California. California consumer privacy act (ccpa). https://oag.ca.gov/privacy/ccpa, 2022 (accessed July 20, 2022).

[36] Matthew Green and Matthew Smith. Developers are not the enemy!: The need for usable security apis. *IEEE Security & Privacy*, 14:40–46, 2016.

[37] Priscilla E Greenwood and Michael S Nikulin. *A guide to chi-squared testing*, volume 280. John Wiley & Sons, 1996.

[38] Sara Gustavsson. An assessment of privacy by design as a stipulation in gdpr. 2020.

[39] Irit Hadar, Tomer Hasson, Oshrat Ayalon, Eran Toch, Michael Birnhack, Sofia Sherman, and Arod Balissa. Privacy by designers: Software developers' privacy mindset. *Journal of Empirical Software Engineering*, 23(1):259–289, February 2018.

[40] Irit Hadar, Tomer Hasson, Oshrat Ayalon, Eran Toch, Michael Birnhack, Sofia Sherman, and Arod Balissa. Privacy by designers: software developers' privacy mindset. *Empirical Software Engineering*, 23(1):259–289, 2018.

[41] Hamza Harkous, Sai Teja Peddinti, Rishabh Khandelwal, Animesh Srivastava, and Nina Taft. Hark: A deep learning system for navigating privacy feedback at scale. In *IEEE Symp. on Security and Privacy*, 2022.

[42] Yongzhong He, Xuejun Yang, Binghui Hu, and Wei Wang. Dynamic privacy leakage analysis of android third-party libraries. *Journal of Information Security and Applications*, 46:259–270, 2019.

[43] US Department Health and Human Services. The Health Insurance Portability and Accountability Act (HIPAA). https://www.hhs.gov/hipaa/index.html, 2024 (accessed Feb 10, 2024).

[44] J. Hoepman. Privacy design strategies (extended abstract). 2014.

[45] J-H Hoepman. Making privacy by design concrete. 2018.

[46] Chris Jay Hoofnagle, Jennifer King, Su Li, and Joseph Turow. How different are young adults from older adults when it comes to information privacy attitudes and policies? *Available at SSRN 1589864*, 2010.

[47] Jaap-Henk Hopeman and Marc van Lieshout. Privacy: a fundamental right.

[48] Stefan Albert Horstmann, Samuel Domiks, Marco Gutfleisch, Mindy Tran, Yasemin Acar, Veelasha Moonsamy, and Alena Naiakshina. "those things are written by lawyers, and programmers are reading that." mapping the communication gap between software developers and privacy experts. *Proc. Priv. Enhancing Technol.*, 2024:151–170, 2024.

[49] Tianjian Huang, Vaishnavi Kaulagi, Mitra Bokaei Hosseini, and Travis Breaux. Mobile application privacy risk assessments from user-authored scenarios. In *Proceedings of the 31st IEEE International Requirements Engineering Conference*, pages 1–12. IEEE, 2023.

[50] International Association of Privacy Professionals. Taxonomy of privacy. https://iapp.org/resources/article/a-taxonomy-of-privacy/a, 2024 (accessed June 1, 2024).

[51] Leonardo Horn Iwaya, Muhammad Ali Babar, and Awais Rashid. Privacy engineering in the wild: Understanding the practitioners' mindset, organisational aspects, and current practices. *IEEE Transactions on Software Engineering*, 2023.

[52] Akshath Jain, David Rodriguez, Jose M del Alamo, and Norman Sadeh. Atlas: Automatically detecting discrepancies between privacy policies and privacy labels. *arXiv preprint arXiv:2306.09247*, 2023.

[53] Vijayanta Jain, Sepideh Ghanavati, Sai Teja Peddinti, and Collin McMillan. Towards fine-grained localization of privacy behaviors. In *IEEE 8th European Symposium on Security and Privacy*, pages 258–277, 2023.

[54] Vijayanta Jain, Sanonda Datta Gupta, Sepideh Ghanavati, and Sai Teja Peddinti. Prigen: Towards automated translation of android applications' code to privacy captions. In *Int. Conference on Research Challenges in Information Science*, pages 142–151. Springer, 2021.

[55] Vijayanta Jain, Sanonda Datta Gupta, Sepideh Ghanavati, Sai Teja Peddinti, and Collin McMillan. Pact: Detecting and classifying privacy behavior of android applications. In *Proc. of the 15th ACM Conf. on Security and Privacy in Wireless and Mobile Networks*, WiSec '22, page 104–118. ACM, 2022.

[56] Harjot Kaur, Sabrina Amft, Daniel Votipka, Yasemin Acar, and Sascha Fahl. Where to recruit for security development studies: Comparing six software developer samples. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 4041–4058, 2022.

[57] Rishabh Khandelwal, Asmit Nayak, Paul Chung, and Kassem Fawaz. Unpacking privacy labels: A measurement and developer perspective on google's data safety section. *arXiv preprint arXiv:2306.08111*, 2023.

[58] William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

[59] Tianshi Li, Elizabeth Louie, Laura Dabbish, and Jason I Hong. How developers talk about personal data and what it means for user privacy: A case study of a developer forum on reddit. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3):1–28, 2021.

[60] Xueqing Liu, Yue Leng, Wei Yang, Wenyu Wang, Chengxiang Zhai, and Tao Xie. A large-scale empirical study on android runtime-permission rationale messages. In *2018 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pages 137–146. IEEE, 2018.

[61] Laura MacLeod, Andreas Bergen, and Margaret-Anne Storey. Documenting and sharing software knowledge using screencasts. *Empirical Software Engineering*, 22:1478–1507, 2017.

[62] European Union Agency For Network and Information Security. Pets controls matrix a systematic approach for assessing online and mobile privacy tools. 2016.

[63] Serge Egelman Noura Alomar and and Jordan L. Fischer. Developers say the darnedest things: Privacy compliance processes followed by developers of child-directed apps. *Proceedings on Privacy Enhancing Technologies*, 2022(4), 2022.

[64] US Department of Education. The Family Educational Rights and Privacy Act (FERPA). https://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html, 2024 (accessed Feb 10, 2024).

[65] Ehimare Okoyomon, Nikita Samarin, Primal Wijesekera, Amit Elazari Bar On, Narseo Vallina-Rodriguez, Irwin Reyes, Álvaro Feal, and Serge Egelman. On the

ridiculousness of notice and consent: Contradictions in app privacy policies. 2019.

[66] United Nations Conference on Trade and Development. Data protection and privacy legislation worldwide. https://tinyurl.com/puev83dt, 2021.

[67] R OpenAI. Gpt-4 technical report. *arXiv*, pages 2303–08774, 2023.

[68] Stefan Palan and Christian Schitter. Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27, 2018.

[69] Rahul Pandita, Xusheng Xiao, Wei Yang, William Enck, and Tao Xie. {WHYPER}: Towards automating risk assessment of mobile applications. In *Presented as part of the 22nd {USENIX} Security Symposium ({USENIX} Security 13)*, pages 527–542, 2013.

[70] Jonathan Parsons, Michael Schrider, Oyebanjo Ogunlela, and Sepideh Ghanavati. Understanding developers privacy concerns through reddit thread analysis. *Joint Proc. of REFSQ-2023 Workshops, Doctoral Symposium, Posters & Tools Track and Journal Early Feedback co-located with the 28th Int. Conf. on Requirements Engineering: Foundation for Software Quality (REFSQ 2023), Barcelona, Catalunya*, 2023.

[71] Zhengyang Qu, Vaibhav Rastogi, Xinyi Zhang, Yan Chen, Tiantian Zhu, and Zhong Chen. Autocog: Measuring the description-to-permission fidelity in android applications. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, pages 1354–1365, 2014.

[72] David Rodriguez, Akshath Jain, Jose M Del Alamo, and Norman Sadeh. Comparing privacy label disclosures of apps published in both the app store and google play stores. In *IEEE European Symp. on Security and Privacy Workshops*, pages 150–157, 2023.

[73] Nayan B. Ruparelia. Software development lifecycle models. *SIGSOFT Softw. Eng. Notes*, 35(3):8–13, 2010.

[74] Raphael Serafini, Marco Gutfleisch, Stefan Albert Horstmann, and Alena Naiakshina. On the recruitment of company developers for security studies: results from a qualitative interview study. In *19th Symposium on Usable Privacy and Security*, pages 321–340, 2023.

[75] J P Shaffer. Multiple hypothesis testing. *Annual Review of Psychology*, 46(1):561–584, 1995.

[76] Yun Shen and Siani Pearson. Privacy enhancing technologies: A review. *Hewlet Packard Development Company. Disponible en https://bit. ly/3cfpAKz*, 2011.

[77] Rocky Slavin, Xiaoyin Wang, Mitra Bokaei Hosseini, James Hester, Ram Krishnan, Jaspreet Bhatia, Travis D Breaux, and Jianwei Niu. Pvdetector: a detector of privacy-policy violations for android apps. In *2016 IEEE/ACM Int. Conf. on Mobile Software Engineering and Systems (MOBILESoft)*, pages 299–300, 2016.

[78] Rocky Slavin, Xiaoyin Wang, Mitra Bokaei Hosseini, James Hester, Ram Krishnan, Jaspreet Bhatia, Travis D Breaux, and Jianwei Niu. Toward a framework for detecting privacy policy violations in android application code. In *Proceedings of the 38th International Conference on Software Engineering*, pages 25–36, 2016.

[79] Daniel J Solove. A taxonomy of privacy. *University of Pennsylvania law review*, pages 477–564, 2006.

[80] Sarah Spiekermann and Lorrie Faith Cranor. Engineering privacy. *IEEE Transactions on Software Engineering*, 35(1):67–82, 2009.

[81] Sarah Spiekermann, Jana Korunovska, and Marc Langheinrich. Inside the organization: Why privacy and security engineering is a challenge for engineers. *Proceedings of the IEEE*, 107(3):600–615, 2018.

[82] Sarah Spiekermann-Hoff, Jana Korunovska, and Marc Langheinrich. Understanding engineers' drivers and impediments for ethical system development: The case of privacy and security engineering. 2018.

[83] Latanya Sweeney. k-anonymity: A model for protecting privacy. *Int. journal of uncertainty, fuzziness and knowledge-based systems*, 10(05):557–570, 2002.

[84] Mohammad Tahaei, Julia Bernd, and Awais Rashid. Privacy, permissions, and the health app ecosystem: A stack overflow exploration. In *Proc. of the 2022 European Symposium on Usable Security*, pages 117–130, 2022.

[85] Mohammad Tahaei, Alisa Frik, and Kami Vaniea. Privacy champions in software teams: Understanding their motivations, strategies, and challenges. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.

[86] Mohammad Tahaei, Adam Jenkins, Kami Vaniea, and Maria Wolters. "i don't know too much about it": On the security mindsets of computer science students. In Thomas Groß and Theo Tryfonas, editors, *Socio-Technical Aspects in Security and Trust*, pages 27–46, Cham, 2021. Springer International Publishing.

[87] Mohammad Tahaei, Tianshi Li, and Kami Vaniea. Understanding privacy-related advice on stack overflow. *Proceedings on Privacy Enhancing Technologies*, 2022(2):114–131, 2022.

[88] Mohammad Tahaei and Kami Vaniea. A survey on developer-centred security. In *2019 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 129–138. IEEE, 2019.

[89] Mohammad Tahaei and Kami Vaniea. "developers are responsible": What ad networks tell developers about privacy. In *Extended Abstracts in CHI Conf. on Human Factors in Computing Systems*, pages 1–11, 2021.

[90] Mohammad Tahaei and Kami Vaniea. Recruiting participants with programming skills: A comparison of four crowdsourcing platforms and a cs student mailing list. In *CHI Conference on Human Factors in Computing Systems*, CHI '22. ACM, 2022.

[91] Mohammad Tahaei, Kami Vaniea, and Naomi Saphra. Understanding privacy-related questions on stack overflow. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–14, 2020.

[92] The Federal Trade Commission. Privacy and security enforcement. 2024 (accessed Feb 10, 2024).

[93] UC - Berkeley - School of Information. Privacy patterns - collaborative development of privacy software design patterns. https://github.com/privacypatterns, 2024 (accessed Feb. 10, 2024).

[94] UC Berkeley - School of Information. Privacy patterns org. https://privacypatterns.org/, 2024 (accessed February 10, 2024).

[95] Varonis. 84 must-know data breach statistics for 2023. https://www.varonis.com/blog/data-breach-statistics, (accessed Feb. 10, 2024).

[96] L. Yu and X. et al. Lou. Can we trust the privacy policies of android apps? In *46th Annual IEEE/IFIP Int. Conf. on (DSN)*, pages 538–549. IEEE, 2016.

[97] Xian Zhan, Lingling Fan, Sen Chen, Feng We, Tianming Liu, Xiapu Luo, and Yang Liu. Atvhunter: Reliable version detection of third-party libraries for vulnerability identification in android applications. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, pages 1695–1707, 2021.

[98] Sebastian Zimmeck, Rafael Goldstein, and David Baraka. Privacyflash pro: Automating privacy policy generation for mobile apps. 2021.

[99] Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel R Reidenberg, N Cameron Russell, and Norman Sadeh. Maps: Scaling privacy compliance analysis to a million apps. *Proc. Priv. Enhancing Tech.*, 2019:66, 2019.

[100] Michael Zimmer. The gaze of the perfect search engine: Google as an infrastructure of dataveillance. In *Web search: Multidisciplinary perspectives*, pages 77–99. Springer, 2008.

A   Survey Questions

Survey questions can be found here: http://tinyurl.com/2p9n49e4

B   Participants' Demographic Information

Table 10 below shows the various demographics of our participants.

C   Details of Solove's Taxonomy

Solove's Taxonomy and the mapping of subcategories.

Table 11: Solove's Categories and Subcategories

Main Category	Solove's Subcategories
Information Collection	Surveillance, Interrogation
Information Processing	Aggregation, Identification, Insecurity, Secondary Use, Exclusion
Information Dissemination	Breach of Confidentiality, Disclosure, Exposure, Increased Accessibility, Blackmail, Appropriation, Distortion
Invasion	Intrusion, Decisional Interference

D   Confidence in Security & Privacy Measures

The hypotheses list for the correlation between confidence in security and privacy measures and various factors are:
– **H1a:** The size of the company correlates with confidence in privacy and security measures.
– **H1b:** The participants' role at the company correlates to confidence in privacy and security measures.
– **H1c:** The education level correlates to confidence in privacy and security measures.
– **H1d:** The presence of a CPO or similar position correlates to confidence in privacy and security measures.

The p-value results of the Chi-Square tests are as follows:

Table 12: P-Value for Hypothesis **H1a** to **H1d**

	H1a	H1b	H1c	H1d
P-Value	0.494	0.654	0.570	0.0007

E   Presence of a CPO or a Similar Role

The participant's knowledge about the presence of a CPO in their company is as follows:

Table 10: Demographic Information about the Participants

Gender	Female (25.48%)	Male (73.41%)	Non-Binary (0.55%)	Other (0.55%)	PnS (0%)
Age	18-25 (19.89%)	26-35 (45.86%)	36-45 (20.99%)	46-55 (8.84%)	>55 (3.87%)
Education	High school (10.22%)	BSc. (61.05%)	MSc. (22.10%)	PhD (1.66%)	Other (3.87%)
Degree	CS/ECE/DS (34.8%)	IT (26.24%)	Business (11.05%)	Other (24.04%)	PnS (3.87%)
Company Size	100+ emp. (50.00%)	50-100 (13.54%)	21-50 (12.43%)	11-20 (7.46%)	0-10 (16.57%)

Table 13: Distribution of Knowledge about a CPO

Yes	No	Unsure	Others
42.6%	38.4%	17.9%	1.1%

The hypotheses list for the correlation between the presence of a CPO/a similar role and the PIA creation, familiarity with PETs, number of privacy breaches, and the company size are:

– H2a: The creation of a PIA correlates to the presence of a CPO or similar position at the company.

– H2b: Familiarity with PETs correlates to the presence of a CPO or similar position at the company.

– H2c: The higher number of privacy breaches correlates to the presence of a CPO or similar position at the company.

– H2d: The size of a company correlates to the presence of a CPO or similar position at the company.

The p-value results of the Chi-Square tests are as follows:

Table 14: P-Value for Hypothesis **H2a** to **H2d**

	H2a	H2b	H2c	H2d
P-Value	0.1005	0.008	0.359	< 0.00001

The distribution of how participants address their compliance questions:

Table 15: Distribution of Sources for Compliance Questions

Lawyer	CPO	Best Practices	Forums	Others
24.2%	25.9%	23.1%	18.5%	8.3%

F   The Creation of a PIA

The hypotheses list for the correlation between the creation of a PIA and the company size and confidence in privacy and security measures are:

– H3a: The size of the company correlates to the PIA creation.

– H3b: The participants' confidence in an organization's privacy and security measures correlates to the PIA creation.

The p-value results of the Chi-Square tests are as follows:

Table 16: P-Value for Hypothesis **H3a** to **H3b**

	H2a	H2b
P-Value	< 0.00001	< 0.00001

The distribution of responses to the creation of a PIA in their company is shown in Figure 5.
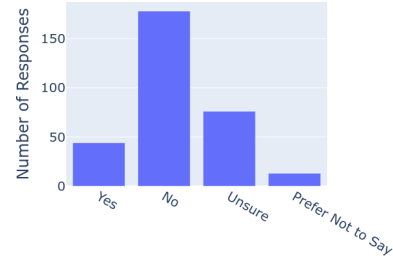
Figure 5: Distribution of Responses to the Creation of a PIA.

G   Privacy by Design Approaches

The distribution of participants who are familiar with PbD:

Table 17: Distribution of Familiarity with PbD Strategies

Role	Yes	No	Unsure	PnS
SD	91 (46%)	54 (27.3%)	49 (24.7%)	4 (2%)

H   Detailed Analysis of PETs' Familiarity

The list of the hypotheses for the correlation between the usage of PETs and the size of the company, participants' confidence, and the presence of the CPO is as follows:

– H4a: The size of the company correlates to the use of PETs.

– H4b: The participant's confidence in an organization's privacy and security measures correlates to the use of PETs.

– H4c: The participant's education level correlates to the use of PETs.

Table 18 shows the results of the hypotheses analysis.

Table 18: P-Value for Hypothesis **H4a** to **H4c**

	H4a	H4b	H4c
P-Value	0.254	0.704	0.529

Table 19: P-Value and H Value for Hypothesis **H5a** to **H5c**

	H5a	H5b	H5c
P-Value	0.04	0.17	0.08
H Value	4.03	7.80	9.83

Table 20: Categories of PETs

Categories of PETs	Definition in Literature [62, 76]	Example from Our Survey
Encryption	A system of communication where the only people who can read the messages are the people communicating.	We use encryption and a number of security features offered by the platform we implement. It is primarily the responsibility of the back-end programmers.
Access Control/ Identity protection	Deals with identifying individuals and controlling access to resources in a system.	We implement role-based access for the various features of our product as well as internally
Anonymity and Pseudonymity	Involves removing personally identifiable information (PII) to prevent individual users from being identified. Pseudonymity involves replacing identifiers with pseudonyms [83].	Data anonymization, our managers would be the primary users for that subject
Differential Privacy	Involves adding noise to the data to protect individual user information while still providing useful insights. It is particularly useful in data analysis and machine learning applications. [25]	We use encryption and al little bit of **differential privacy** where it is applicable and it varies from project to project with who is tasked with implementing these features.
Secure Communication/ VPN	Involves encrypting all communications within the software using standard protocols like HTTPS and SSL/TLS.	All of our internal communication is done over an internal VPN, and all web access is done with https.
Privacy-Enhance Anti Web Tracking	Involves blocking attempts of different types of trackers to monitor users' online activity and personal data.	-

I   Factors Influencing Usage of Forums

To further evaluate the impact of the size of the company, familiarity with PETs, and the presence of a CPO on the usage of developer forums, we employed the Kruskal-Wallis test which is a non-parametric test that is used to compare two or more independent samples for statistically significant differences between groups [58]. Below is the list of hypotheses for the frequency of the usage of the developers' forums:

– H5a: The size of the company correlates to the use of developer forums to ask privacy-related questions.

– H5b: The presence of a Chief Privacy Officer or similar position at a participant's organization correlates to the use of developer forums to ask privacy-related questions.

– H5c: Familiarity with PETs correlates to the use of developer forums to ask privacy-related questions.

As shown in Table 19 when comparing forum usage with the size of the company, a statistically significant difference was found between the groups ($H - Value = 4.03, p - value = 0.04$). However, no significant difference was noted when comparing forum usage with the presence of a Chief Privacy Officer (CPO) ($H - value = 9.83, p - value = 0.08$) or with the usage of Privacy Enhancing Technologies (PETs) ($H - value = 3.92, p - value = 0.56$). These findings suggest that only the size of the company is more likely to influence the frequency with which developers consult forums for privacy-related inquiries.

J   Details for the Location Analysis

Below is the list of hypotheses for location analysis.

– H6a: The participants' confidence in their organization's privacy and security measures correlates to their region of origin.

– H6b: The presence of a CPO or similar position at a participant's organization correlates to their region of origin.

– H6c: The participants' creation of a PIA correlates to their region of origin.

– H6d: The participants' organization being a victim of a breach of privacy correlates to their region of origin.

– H6e: The participants' familiarity with PbD strategies correlates to their region of origin.

– H6f: The participants' use of PETs correlates to their region of origin.

– H6g: The participants' familiarity with the CCPA correlates to their region of origin.

– H6h: The participants' familiarity with the GDPR correlates to their region of origin.

K   Qualitative Analysis Guidelines

Table 20 shows the different categories of PETs and Table 21 describes the privacy taxonomy, both of which were considered as guidelines for our qualitative analysis.

Table 21: Taxonomy of Privacy

Taxonomy of Privacy	Solove's Definition [79]	Example from IAPP [50]	Example from Our Survey
Surveillance	Watching, listening to, or recording of an individual's activities	A website monitoring the cursor movements of a visitor while visiting the website.	Privacy is the ability to keep information or activities out of public knowledge
Interrogation	Questioning or probing for personal information	An interviewer asking an inappropriate question, such as marital status, during a employment interview.	As far as I'm the internet, not asking for private information from our customers such as addresses or any sensitive information.
Aggregation	Combining of various pieces of personal information	A credit bureau combining an individual's payment history from multiple creditors.	Keeping unnecessary information from being exchanged at the minimum amount possible.
Insecurity	Carelessness in protecting information from leaks or improper access	An e-commerce website allowing others to view an individual's purchase history by changing the URL (e.g. enterprivacy.com?id=123)	Having confidential and private information secured and stored away safely from malicious users.
Identification	Linking of information to a particular Individual.	A researcher linking medical files to the Governor of a state using only date of birth, zip code and gender.	I think it can be defined as a set of personal information of each individual that should not be accessible to other people
Secondary Use	Using personal information for a purpose other than the purpose or which is was collected	The U.S. Government uses census data collected for the purpose of apportioning Congressional districts to identify and intern those of Japanese descent in WWII.	Ensuring the minimum amount of data is available only to those that genuinely need it for business purposes, and that it's only available for the specified amount of time that the data is needed.
Exclusion	Failing to let an individual know about the information that others have about them and participate in its handling or use	A company using customer call history, without the customer's knowledge, to shift their order in a queue (i.e. "Your call will be answer in the order [NOT] received")	to have the authority of controlling information about yourself who can or can not see. to be from from any interference, and to be able to interact with anyone I want.
Breach of Confidentiality	Breaking a promise to keep a person's information confidential	A doctor revealing patient information to friends on a social media website.	Having confidential and private information secured and stored away safely from malicious users.
Disclosure	Revealing truthful personal information about a person that impacts the ways others judge their character or their security	A government agency revealing an individual's address to a stalker, resulting in the individual's murder.	Data must be kept safe, and users need that information to be seen only by those they authorize.
Exposure	Revealing an individual's nudity, grief, or bodily functions	A store forcing a customer to remove clothing revealing a colostomy bag.	Freedom of your own information.
Increased Accessibility	Amplifying the accessibility of personal information	A court making proceeding searchable on the Internet without redacting personal information.	A state where one can be sure no one else knows what they are doing
Blackmail	Threatening to disclose personal information	A dating service for adulters charging customers to delete their accounts.	-
Appropriation	Using an individual's identity to serve the aims and interests of another	A social media site using customer's images in advertising	Being able to be secure in your information so that none of it gets accessed or leaked by outside sources
Distortion	Disseminating false or misleading information about an individual	A creditor reporting a paid bill as unpaid to a credit bureau.	Privacy refers to an individual's right to control [..]on. This includes protecting sensitive data from [..], and providing individuals with the ability to access, correct, or delete their PI.
Intrusion	Disturbing an individual's tranquility or solitude	An augmented reality game directing players onto private residential property.	The right to be let alone,or freedom from interference or intrusion.
Decisional Inference	Intruding into an individual's decision regarding their private affairs	A payment processor declining transactions for contraceptives	The right to be let alone,or freedom from interference or intrusion.

Privacy Communication Patterns for Domestic Robots

Maximiliane Windl[1,2], Jan Leusmann[1], Albrecht Schmidt[1,2], Sebastian S. Feger[1,3], Sven Mayer[1,2]

[1] *LMU Munich, Germany*
[2] *Munich Center for Machine Learning (MCML), Germany*
[3] *Rosenheim Technical University of Applied Sciences, Germany*

Abstract

Future domestic robots will become integral parts of our homes. They will have various sensors that continuously collect data and varying locomotion and interaction capabilities, enabling them to access all rooms and physically manipulate the environment. This raises many privacy concerns. We investigate how such concerns can be mitigated, using all possibilities enabled by the robot's novel locomotion and interaction abilities. First, we found that privacy concerns increase with advanced locomotion and interaction capabilities through an online survey ($N = 90$). Second, we conducted three focus groups ($N = 22$) to construct 86 patterns to communicate the states of microphones, cameras, and the internet connectivity of domestic robots. Lastly, we conducted a large-scale online survey ($N = 1720$) to understand which patterns perform best regarding trust, privacy, understandability, notification qualities, and user preference. Our final set of communication patterns will guide developers and researchers to ensure a privacy-preserving future with domestic robots.

1 Introduction

Smart assistants have long become integral parts of many homes, as they make life more enjoyable by providing entertainment or supporting with daily chores. Most of these devices are either placed in a dedicated area, such as smart speakers or have minimal interaction capabilities, such as robot vacuums. Despite their restricted movement and interaction, they already cause various privacy concerns [26, 27, 50] as their sensors collect and process sensitive data. Such concerns include the smart assistant transmitting data without explicit consent [26] or being exposed to microphones that are always listening and sharing recordings with third parties [27]. However, through advancements in AI and robotics,

future smart assistants will not remain static and passive (c.f., Amazon Astro). Quite the contrary – they will gain various locomotion and interaction capabilities, allowing them to enter all areas and even physically manipulate the environment. Such domestic robots will increase our convenience as they take over tasks like folding laundry or cleaning bathrooms. However, this will make them even more intrusive as the robots can access all rooms or even search through personal belongings, paving the way for various privacy concerns.

Due to their advanced locomotion and interaction capabilities and potential for social bonding, domestic robots pose completely new threats to users' psychological, social, and physical privacy [32]. Users, for example, report being concerned about getting accidentally recorded while the robot moves past or interacts with other entities [28]. Moreover, humanoid robots pose a particular threat to users' privacy, as they provoke trust, leading to users' willingly sharing feelings and sensitive information [48]. Further, their humanoid appearance lets people underestimate their capabilities as they relate them to human capabilities [28]. As a result, experts demand that robots regularly communicate their privacy states to users, such as unambiguously indicating whether they are currently recording [24]. Even though there have been suggestions for such communication patterns [32], research is scarce and lacks an encompassing picture. Thus, we do not know which patterns evoke trust, are understandable, have good notification qualities, and are favored by users.

To close this gap, we first investigated the impact of locomotion and interaction capabilities on privacy concerns. Then, we investigated how domestic robots can communicate their sensor states to allow users to assess potential privacy risks. We explore two dimensions that contribute to privacy risks: a) the locomotion (4 levels) and b) interaction (3 levels) capabilities. We conducted an online survey (N=90) to understand how the resulting $4 \times 3 = 12$ scenarios affect user privacy concerns and investigated reasons for concerns. We then elicited communication patterns in three focus groups (N=22) that allow users to assess the robot's sensor states (cameras, microphones, and network connectivity). Finally, we conducted

a large-scale survey (N=1720) to understand which patterns performed best regarding trust, privacy, understandability, notification qualities, and general user preference.

This paper provides a path to allow domestic robots to enter our homes while keeping privacy concerns low. First, we found that advanced locomotion and interaction capabilities increase users' concerns. Second, we provide a set of 86 communication patterns to indicate the robots' microphone, camera, and connectivity states. Finally, we found that most of our elicited communication patterns scored equally well, showing that which pattern to use depends on the characteristics of the situation. To the best of our knowledge, this paper is the first to provide (1) an understanding of how increased locomotion and interaction capabilities of future smart assistants affect users' privacy concerns, (2) construct an encompassing set of various communication patterns for domestic robots to indicate the state of their privacy-relevant capabilities, and (3) provide insights into the quality of the communication patterns. Furthermore, we developed an interactive web application to facilitate the exploration, filtering, and retrieval of appropriate communication patterns based on designers' and researchers' diverse needs and preferences. With this, our set of patterns will guide developers and researchers in ensuring a privacy-preserving future with domestic robots.

2 Related Work

First, we report on privacy in smart home contexts: The specific risks, users' concerns, and mitigation strategies. Second, we highlight work on privacy concerns of domestic robots.

2.1 Privacy in Smart Homes

Through their placement in our intimate spaces, smart home devices are exceptionally prone to revealing sensitive information when exploited. Research, for example, showed how data from smart devices allows retracing identities [42], tracking user behavior [4], revealing the number of people in a household, or their sleeping and eating routines [40].

While some users are unable to pinpoint the concrete dangers posed by smart devices [20, 34, 35], they still feel a sense of unease or have concrete privacy concerns when in their vicinity [50]. Such concerns include personal data being revealed without explicit consent [26], for example, through always-listening smart speakers that share these data with third parties [27]. Prior research also found a diverging danger perception regarding different sensor types [50]. Users are most concerned about cameras and microphones [12, 50] and mostly consider temperature or motion sensors [50] less concerning. Some even express clear skepticism that these sensors cause any concern at all [9, 13, 58].

Prior research also investigated approaches to counter these concerns, including technological measures, such as implementing traffic shaping techniques [5], auto-configuring smart

devices and implementing automatic updates [30], or introducing frameworks that automatically adjust the privacy level in smart homes depending on contexts [41] or pre-defined privacy zones [7]. Moreover, through co-design studies, Yao et al. [55] suggest different control mechanisms, such as disconnecting devices from the internet and keeping data local, increasing transparency and control, and providing access control through different modes. Next to these approaches, a more recent thread of research focuses on tangible control mechanisms [3, 14, 38, 52]. A major advantage of these mechanisms is their high understandability, which instills trust and guarantees inclusivity, especially for people with low technological understanding [3, 52]. Moreover, Chalhoub et al. [12] found that physical camera shutters are especially desired in privacy-sensitive locations, such as bathrooms.

Sensitive data collected in homes can be exploited, raising various privacy concerns. Yet, traditionally, smart devices were static and had limited interaction capabilities. Future smart assistants will have advanced capabilities through advancements in AI and robotics, enabling completely new ways to invade privacy. Hence, we must understand how such increased capabilities affect users' privacy in home contexts.

2.2 Privacy and Domestic Robots

Domestic robots have advanced locomotion and interaction capabilities, enabling them to access all private spaces. This means that their presence might affect not only informational privacy but also physical, psychological, and social privacy [32]. Many domestic robots are, for example, equipped with mobile cameras, enabling them to take images of users or even children in locations such as the bedroom and bathroom, collect spatial information, or witness conversations unnoticed by the users [10, 15, 46]. Moreover, their verbal communication abilities, often paired with a humanoid appearance, lead to people deliberately sharing sensitive information [32, 48].

Even though prior research emphasized the dangers caused by the robots' mobility and physicality [11], users are more concerned about the institutional aspects of their privacy [31], such as how manufacturers handle their data and tended to underestimate the impact of domestic robots on their physical privacy. Yet, users report concerns about the robot being misused for malicious purposes, such as stalking or hacking [31]. Moreover, users in an interview study by Lee et al. [28] reported not being concerned about the robot recording their interactions as long as they were aware of it. However, the interviewees were concerned about accidental recordings that might happen while the robot moves or interacts with other entities. Overall, participants agreed they wanted to be notified about such accidental recordings. The authors also found that participants underestimated the robot's capabilities due to its humanoid shape, which led them to believe that the camera was functioning like human eyes and could not see objects behind its back. Hence, they conclude that users must be

thoroughly informed about the robots' exact capabilities [28].

Experts demand that robots actively communicate when they surveil specific areas [32]. Especially only giving a one-time notice upon purchase is not enough; Instead, robots should give dynamic feedback to regularly communicate their privacy state to users [24]. Lutz et al. [32] conducted expert interviews to elicit privacy mitigation strategies for robots. Their approaches include being able to switch off a robot, limiting its movement space, employing data anonymization, or even designing the robot's humanoid features (i.e., its eyes and ears) in a way to signal if data is being collected.

Domestic robots raise various novel privacy concerns. Thus, experts demand that they regularly communicate their privacy states. Yet, we currently lack a systematic understanding of what communication patterns domestic robots can employ and we do not know which patterns perform best regarding measurements such as understandability and trust.

2.3 Research Questions

We investigate how locomotion and interaction influence users' privacy concerns and how future domestic robots can effectively communicate the state of their privacy-relevant capabilities through the following three research questions:

RQ1. Prior research showed that current smart devices cause various privacy risks [4, 40, 42], making users concerned about their privacy [12, 26, 27, 50]. Yet, current smart home devices are static or have limited interaction capabilities. In contrast, future domestic robots will have increased capabilities, making them even more invasive. Prior research already showed that domestic robots introduce a new range of risks and concerns [11, 28, 46], yet we do not know how the different levels of interaction and locomotion capabilities impact user concerns. Therefore, we ask in our first research question **(RQ1): How do privacy concerns change with increasing levels of locomotion and interaction capabilities?**

RQ2. Prior research points to the additional risks posed by domestic robots, such as being able to follow us around [46], enter all areas [11], or even make accidental recordings [28]. In response, experts call for domestic robots to communicate their privacy-relevant states to the user regularly [24, 32]. However, research in this regard is scarce. Hence, we ask in our second research question **(RQ2): Which patterns should domestic robots employ to communicate their privacy-relevant functionalities to users?**

RQ3. Finally, we need to find out which patterns perform best. In detail, we want to find out which patterns users trust most, which they felt to increase their privacy, which they found most understandable, which they believed to have the best notification qualities, and which they would prefer their smart assistant to use. Hence, we ask in our last research question **(RQ3): Which communication patterns perform best regarding trust, privacy, understandability, notification qualities, and general user preference?**

3 Study I: Locomotion and Interaction Impact

We first set out to understand how increased locomotion and interaction capabilities influence users' privacy concerns in the context of domestic robots. While prior work points to the risks introduced by domestic robots' increased capabilities [11, 28, 46], research on users' concrete concerns is scarce or even shows that users underestimate the impact of robots on their physical privacy [31]. Hence, we conducted an online survey using Prolific to answer our first research question (**RQ1**). We acquired ethics approval for the survey.

3.1 Survey Construction

As prior work showed that a multitude of different factors, such as the sensors [35, 50], device manufacturers [36, 56], perceived device utility [56], and familiarity [6, 50] influence users' privacy concerns, we focused on creating descriptions for the various smart assistants with as few biasing factors as possible. Therefore, we used sole textual descriptions and refrained from using pictures or illustrations to not create associations with existing smart home devices or specific manufacturers; relying solely on text is an approach also followed by related work when capturing perceptions of future scenarios [49]. Furthermore, we aligned all texts and only varied the locomotion and interaction capabilities descriptions. Four researchers, two with expertise in privacy and two in human-robot interaction, collaboratively created the different interaction and locomotion stages by clustering the most popular smart assistants according to their capabilities and extending them with the full human-like capabilities, *world movement* and *full interaction* to represent future smart assistants. This process resulted in three interaction stages and four locomotion stages, which we combined to create descriptions for 12 smart assistants. All descriptions used the following structure: `"Imagine the following scenario – You own a smart assistant that you are using in your home. It has the following capabilities: [Locomotion Capability] + The smart assistant possesses sensing abilities that enable it to comprehend its surroundings + [Interaction Capability]."` We revised these textual descriptions through several rounds of discussions before we conducted pilot tests with two researchers in the field of human-computer interaction who were not involved in this project and with 10 participants from Prolific. In response to piloting, we made the locomotion capability descriptions more comprehensive. This resulted in the following texts:

Locomotion Capabilities. *Stationary:* The smart assistant is stationary, which means it is constrained to the exact position where you placed it. *Linear Movement:* The smart assistant can move along a defined path, meaning its movement is constrained by the path you defined. *Planar Movement:* The smart assistant can move freely around flat, even surfaces,

which means that it can freely move around all accessible areas as long as they are on the same floor. *World Movement:* The smart assistant can move freely across all areas, which means it can move around all accessible areas, even if they are not on the same floor.

Interaction Capabilities. *Passive Interaction:* Yet, the smart assistant can not physically manipulate the environment, objects, or itself. This implies it can perceive individuals and objects within its field of view and analyze associated information. *Limited Interaction:* While the smart assistant can automatically adjust its orientation to observe its full surroundings, it can not physically manipulate the environment or objects. This implies it can perceive individuals and objects and analyze associated information. *Full Interaction:* The smart assistant can automatically adjust its orientation to observe its full surroundings and physically manipulate the environment, objects, and itself. This implies it can perceive individuals and objects and analyze associated information.

We started the survey with demographic questions, used the IUIPC questionnaire [33] to understand participants' general privacy perception, and the ATI questionnaire [18] to understand the sample's technical affinity. Afterward, we confronted participants with all 12 smart assistants in random order. After each smart assistant, we asked the participant to respond to *"I am strongly concerned about my privacy due to the presence of the smart assistant"* on a 100-point slider ranging from strongly disagree to strongly agree. We used a visual analog scale (VAS) without ticks to prevent the responses from converging around the ticks, cf. [37]. Moreover, we decided to use VAS, as they have been shown to lead to more precise responses and higher data quality [19]. Finally, as VAS collect continuous data, they allow for more statistical tests [43]. In line with recommendations for scale development, we phrased the statements strongly as mildly phrased statements have shown to result in too much agreement [16].

Additionally, we asked participants to explain their ratings using free text. To ensure the quality of our data, we saved a timestamp after each section and used an attention check item that randomly asked to either set a slider all the way to the right or the left. For the full questionnaire, see Sec. A.1.

3.2 Participants

We recruited 151 participants via Prolific. We did not use any reputational filters, and our sample had a mean of 337 approved tasks ($SD = 292$). We had to exclude 61 participants for (1) giving low-effort responses (N=48), meaning they explained their ratings with only 2-4 words (e.g., "NA," "i trust") or copied the same response in all 12 conditions, (2) straight-lining, i.e., consistently rating all conditions with 0 or 100 (N=9), (3) failing our attention check (see Sec. A.1, question 4c) (N=2), (4) entering mismatched demographics between Prolific and our survey (N=1), and (5) completing the survey three standard deviations faster than the mean (N=1).
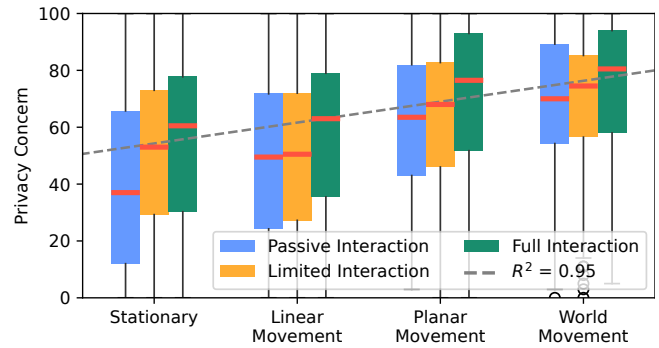
Figure 1: Participants' mean privacy concern over all locomotion and interaction capabilities with boxplots. The trendline represents the change in relation to the locomotion capability.

The final 90 participants (47 male, 42 female, and one preferred not to disclose) were between 19 and 62 years old ($M = 32.9, SD = 9.75$). They were located on three continents (Europe, America, and Africa). Most participants (8) lived in Poland, the United Kingdom, and Italy, followed by Spain (7), South Africa (7), and Portugal (6). Among the participants, 72 were employed full-time, 13 were employed part-time, and five were not in paid work. Moreover, 17 participants were students. Our participants' technical affinity according to the ATI scale [18] was 4.1 ($SD = 0.8$) measured on a 6-point scale. We employed the IUIPC questionnaire [33] using a 7-point Likert scale to understand their general perception of privacy. The results revealed an average rating of 6.2 ($SD = .9$) for Awareness, 5.6 ($SD = 1.1$) for Control, and 5.5 ($SD = 1.1$) for Collection. These scores indicate a relatively high level of privacy concerns, cf. [22]. The survey took ∼16min, and they were compensated with 2.40£.

3.3 Data Analysis

We used Python and R to analyze our quantitative data and affinity diagramming [21] for the qualitative data. Here, we printed all statements so two researchers could collaboratively extract the themes by grouping them. We then created headers for each group, frequently rearranged the items, and refined the themes through multiple discussion rounds.

3.4 Quantitative Results

As our data were not normally distributed ($W = .944, p < .001$), we used an ART ANOVA [54], which revealed significant effects for LOCOMOTION ($p < .001$)($\eta_p^2 = .226$) and INTERACTION ($p < .001$) ($\eta_p^2 = .059$) while indicating no interaction effect ($p > .4$), see Fig. 1. Pairwise post hoc tests using Wilcoxon signed rank tests with Holm–Bonferroni corrections applied showed that the LOCOMOTIONS are rated significantly different (*linear × stationary $p < .05$*, and all others $p < .001$). Moreover, all INTERACTIONS were rated

significantly different (*passive* × *limited* $p = .004$, and all others for all $p < .001$). We assumed an equidistant distribution between the smart assistants and fitted a line to all mean concern ratings, see Fig. 1. As all trendlines are positive, we conclude that higher locomotion freedom and more interaction capabilities lead to greater privacy concerns.

3.5 Qualitative Results

From the free text descriptions of the participants, we formulated three themes: *Concerns Rooted in Locomotion*, *Concerns Rooted in Interaction*, and *Additional User Concerns*.

3.5.1 Concerns Rooted in Locomotion

We report our participants' explanations of how the different LOCOMOTION capabilities influence their privacy concerns.

Stationary. Our participants felt most in control over what the assistant could hear and see in the *stationary* condition. P31, for example, explains that they *"would try to place it in a space where no personal activities or situations [are] accessible."* Such a non-concerning space could be the kitchen, where the participants do not expect personal conversations to occur but consider the smart assistant especially useful for playing music or providing recipes (P43).

Linear. Our participants explained that the *linear* movement would reduce their concerns as they can specify the areas the assistant can access. P53, for example, states: *"Because the path is pre-defined, [...] I'd simply avoid putting the smart assistant in the rooms I would like to have privacy in."* P29 further states that they would redefine the assistant's path should their preferences or concerns change.

Planar. In contrast to the two more restricted movement capabilities, the *planar* movement increased our participants' privacy concerns significantly, as P14 explains: *"If the assistant is left to roam free, it can collect information at will, and that is a clear security and privacy concern."* Yet, participants still felt the assistant's inability to climb stairs or move to different floors helped in preserving some privacy: *"Due to its limitation to one floor I might feel a bit safer with my privacy, I can move downstairs or upstairs"* (P43).

World. Our participants were most concerned in the *world* movement condition as they feared the smart assistant could follow them everywhere, leaving no protected space: *"Being able to move even to different floors means there is no safe place in the house"* (P83). Moreover, participants were concerned about the assistant showing up unexpected (P56): *"It's hard to avoid it popping up unexpectedly, isn't it?"*

3.5.2 Concerns Rooted in Interaction

We now report the influence of the different INTERACTION capabilities on participants' privacy concerns.

Passive Interaction. In the *passive* condition, most participant responses again revolved around the notion of control.

Participants felt less concerned about their privacy, as they would have *"full control on what it sees"* (P66), and P70 mentioned that the assistant could *" only see what I want."* Here, familiarity also played a role as participants knew stationary smart assistants from their daily life, as P83 states: *"That's the standard setup of intelligent assistant, so no concern."*

Limited Interaction. In contrast, the *limited* interaction capability made our participants way more concerned. Here, P27, for example, compared such a smart assistant to a big brother's eye that would follow them around. In addition, due to its new capabilities, our participants felt less in control over what the smart assistant could perceive: *"It can adjust its sight to some parts I do not want to"* (P75).

Full Interaction. In addition to the concerns reported regarding the *passive* and *limited* interaction capabilities, our participants were now also concerned about the assistant entering all spaces, leaving virtually no room for privacy. As the robot could now *"probably open doors and enter areas in times where [I] don't want [it] to"* (P43). Additionally to this concern, participants also reported a sense of unease thinking about how the assistant could physically *"search the data it wants"* (P1) by searching through personal belongings (P30).

3.5.3 Additional User Concerns

Our participants also reported additional concerns not rooted in the robot's interaction and locomotion capabilities. The smart assistant's internet *connectivity* was the most commonly mentioned concern ($N = 22$). Here, participants were concerned that the smart device might share their data, either with the device manufacturer or third parties. For example, P17 stated that they are *"always concerned about the type of data [smart devices] can provide to their creator"* and P46 said that they would *"question if the assistant passes what it perceives to a third party or a remote server."* As a possible remedy, P43 suggested having an offline assistant or one that can only connect to specific applications. The second most common ($N = 20$) concern was the assistant's video *camera* sensor, as P57 stated: *"I don't like to be watched."* P1 was especially concerned about being filmed in intimate situations: *"They can probably see me naked while I leave the bathroom."* This concern was followed by the *audio* sensor, which 11 participants mentioned. P27, for example, was concerned that the assistant *"might be recording conversations"*, and P43 mentioned that they would even be concerned about the stationary assistant having good enough microphones to eavesdrop on conversations that might be happening in a different room. Moreover, ten participants mentioned being concerned about the *assistant getting hacked*, giving criminals access to their sensitive data. P52, for example, wrote: *"Someone could hack onto it and know how my home is "built" and break into it."* Finally, eight participants were concerned about the assistant *storing data*: *"I do not know where the data is saved"* (P69). Less commonly mentioned were concerns regarding the *de-*

tection of activity data ($N = 5$) and *identification* ($N = 1$).

We focus the remainder of this paper on clearly communicating the state of the capabilities our participants most frequently mentioned: internet connectivity, cameras, and audio sensors. Yet, it is important to note that concerns go beyond the pure collection of data, e.g., what could be inferred from the collected data. Yet, to clearly define the scope of this paper, we leave such investigations to future work.

4 Study II: Eliciting Communication Patterns

While prior research demanded that domestic robots clearly communicate their current privacy state to users [24, 32], research on concrete communication patterns is lacking. Hence, we ran three focus groups with 22 participants to answer (**RQ2**). We used focus groups to join diverse perspectives and spark creativity. Our ethics committee approved the study.

4.1 Procedure

We asked participants for their informed consent and demographics. We continued with an introductory round and prior experiences with smart homes and robotic systems. Next, we presented a variety of smart home assistants using pictures and short video clips, aiming to portray the diverse landscape of capabilities and shapes. We started with stationary devices without interaction capabilities and ended with humanoid robots with world movement and full interaction capabilities. As most participants had little experience with robotic systems, it was important to show the diversity to elicit a set of patterns applicable to various domestic robots. Next, we focused on the sensing capabilities of domestic robots, ensuring that they knew that the robots were not restricted to a camera and microphone placed visibly in the front but that the sensing units could be placed everywhere. We then split them into pairs to discuss the risks introduced by domestic robots.

Next, we presented two privacy-relevant future scenarios with domestic robots. In the first scenario, a person sat at the kitchen table, reviewing medical files while discussing the results with their doctor. In the second scenario, a person was getting ready in the bathroom while ranting about their day. We included a domestic robot in both scenarios to make clear that there are scenarios where robots can help with chores but where we also require privacy. Next, we discussed how current smart assistants communicate their privacy state, showing the Alexa Show's camera shutter and the Amazon Echo's microphone-mute button. We contrasted this with how humans communicate that they are not listening or watching.

We introduced the four locomotion stages and the three interaction capabilities. We divided them into pairs and did three rounds of discussions and presentations. For each round, every pair had the same interaction capability: passive interaction, limited interaction, or full interaction. Yet, every pair

had a different locomotion capability to join diverse perspectives and animate them to consider their robot's specific skills. We had at least two physical variants of each locomotion and interaction capability in the room to have something graspable for them to interact with. We randomized the order of the interaction capabilities for each focus group to reduce biases. We handed them pen and paper to sketch their ideas. Examples of the sketches can be found in the Appendix Fig. 4. The task was to develop as many communication patterns as possible that signify the state of the camera, microphone, and internet connectivity. We focused on cameras, microphones, and internet connectivity as we found that users were most concerned about them in our first survey. Finally, we had a last group discussion to reflect on the communication patterns invented and to discuss the future of domestic robots in general.

4.2 Participants

We recruited 22 participants (12 male, and 10 female) based on demographics they provided through a pre-screening questionnaire via a university mailing list. They were between 19 and 65 years old ($M = 29.3$, $SD = 11.4$) with different cultural and educational backgrounds, and came from eight different countries, namely Germany (8), India (5), USA (3), China (2), Brazil (1), South Korea (1), Jordan (1), and Bangladesh (1). They also had different educational backgrounds in computer science (6), biology (3), physics (3), electrical engineering (2), psychology (2), mathematics (2), data science (1), journalism (1), political science (1), and business (1). Their average technical affinity according to the ATI scale [18] was 4.1 ($SD = 0.9$). Six participants had never interacted with a robotic system before, nine 1-3 times, one 4-7 times, and six more than 7 times. They received 20€ for the 2h session.

4.3 Results

We transcribed all focus groups and analyzed the data using thematic analysis [8] and Atlas.ti. First, two researchers independently open-coded the data. They then discussed their codes, resolved ambiguities, and formed code groups. Afterward, a third researcher joined to refine the code groups and extract overarching themes. This process resulted in 202 individual codes, 15 code groups, and six themes. The themes INTERVENTIONS and AWARENESS MECHANISMS form our 86 communication patterns. We also identified the themes TRUST and USABILITY, classifying our patterns further and discussing their applicability. The last theme is HUMANOID VS. NON-HUMANOID, discussing anthropomorphic robots.

4.3.1 Interventions

This theme consists of all communication patterns that not only signal that a capability is deactivated but physically

prevent its function. The patterns in this theme can be further divided into *physical robot constraints*, *physical location constraints*, and *attached props control*. *Physical robot constraints* describes all communication patterns where the robot physically interferes with its capabilities. It ranges from less extreme interventions, such as turning the sensors away (P2, P5, P7, P9, P15), covering the ears with the hands (P8, P20), or detaching individual sensors (P2, P9, P12, P13, P15, P17), to extreme interventions, such as removing the whole head (P2, P16) or even self-destruction (P10, P12). P13 explains how detaching the sensors could look like: *"Having a camera, microphone and a connectivity module and using the hands; basically, the robot taking it off itself, making it very clear that it's not connected."* In *physical location constraints*, our participants discussed interventions that restrict the robot's movement and, thus, its functionalities. Such patterns included the robot blocking its own movement (P2, P5, P10, P12), going to its docking station (P5, P6, P7), or even entering physical confinement (P2, P5, P15, P12, P16, P20, P19), as P15 explains: *"[...] a box, like a parking spot, which is like a Faraday box, where no Wi-Fi connection can come through. It's a non-transparent box, and it's soundproof."* The last group, *attached props control*, contains all patterns where the robot has a privacy prop attached, which blocks the robot's functionality. Here, our participants referred to classical camera shutters (P2, P21) but also cables (P5, P11) and switches (P4, P5, P14) that are solely attached to physically interfere with a capability *"and when you want to shut it down, just press the switch like a light, and everything will be shut off"* (P14).

4.3.2 Awareness Mechanisms

In contrast to the above theme INTERVENTIONS, AWARENESS MECHANISMS do not physically prevent a capability but raise users' awareness of the robot's current privacy state. This theme consists of the following code groups: *Physical robot manipulation*, *attached props feedback*, *environment interaction*, *visual feedback*, and *audio feedback*. *Physical robot manipulation* contains all the ways a robot can change its own appearance to indicate its current privacy-relevant state, including using hand gestures, such as covering the eyes to signal that it is not watching or crossing the arms to signal the Wifi is disconnected (P20, P22), as P20 explains *"you cross your arms out of frustration."* Other suggestions included showing empty connectivity ports to the user (P16), retracting sensors (P19), and signaling disengagement through the body posture (P5, P8, P12, P16, P19, P22): *"These robots could also just let the arms fall, you can see that the motors and everything are disengaged"* (P12). Lastly, the participants also suggested that the robot changes its shape to signal that its capabilities are not activated (P1, P2, P5, 19): *"So it could be in a special form when it's active, but while it's deactivated, it could fall into a different form so you know... shape changing"* (P19). The group *attached props feedback* encompasses

all patterns where the robot has privacy-specific artifacts attached to communicate the privacy state. This included waving a banner to signal that a capability was deactivated (P7), or attaching a light band (P5), or fake antenna: *"Put an antenna or something physical on there that has no use except that it would maybe illuminate red if it's not connected to the internet"* (P20). In *environment interaction*, our participants discussed how the robot could use smart lights installed in the home to communicate its privacy state (P2, P7): *"I see a flickering of the light; So it indicates, okay, it's not listening anymore"* (P2). In *visual feedback*, we summarized all traditional patterns requiring a screen or using simple light feedback (P1, P2, P5, P7, P8, P11, P15, P20 - P22). Our participants had diverse ideas of what could be displayed on the screen, ranging from simple text (P7, P8) to symbols (P20, P22), gestures (P3), and a humanoid face (P7, P20, P22) to turning the screen off (P2). Lastly, our participants suggested using some form of audio feedback, such as playing distinct sounds (P4, P10, P20, P22) or using the robot's voice (P7, P8, P20, P22): *"It says: I'm not listening now"* (P20).

4.3.3 Trust

This theme describes the factors influencing trust in communication patterns. Here, participants discussed that the type of robot determines their preferred communication patterns. While they considered stationary robots as not very invasive and, thus, requiring less invasive strategies (P5), they discussed that robots with more extreme capabilities also require extreme interventions (P10, P11): *"I think that self-destruct is still useful. When your robot has so many capabilities, you also need very strong limitations"* (P11). Our participants also discussed that they prefer manual over system control for such invasive robotic systems. That means they preferred mechanisms where the robot can not reactivate its functionalities by itself (P2, P10, P15, P16, P22). P15 suggested hiding the detached sensors from the robot or adding a physical lock so the robot can not free itself: *"We thought about a lock from the outside so the robot could close the lid by itself, but then the human could have like a mechanical lock that he or she puts from the outside to be sure that the robot itself can't reopen it."* Lastly, our participants also discussed how AWARENESS MECHANISMS require more trust in the robot and its manufacturer than INTERVENTIONS (P5, P10, P11, P13, P15, P16): *"It obviously requires some trust in the company that the lights actually state the true status of the device"* (P15). In contrast, P13 explained what they like about INTERVENTIONS: *"Even if we can't really trust the company – it's a physical barrier."*

4.3.4 Usability

Our participants discussed how the situation influences the applicability of the different patterns and how familiarity, intuitiveness, and joy of use affect their perception of the patterns.

Our participants discussed, for example, that audio feedback is most effective when the robot is not in the same room or hidden somewhere (P1, P2, P10): *"It should also give some audio feedback. So if it's somewhere under my couch, and I can't see it, I know if it's on or off"* (P10). Besides, our participants also discussed that many of the INTERVENTIONS are unsuitable if the robot is currently doing a task (P1, P15, P20): *"If you tell it: Just go away! That doesn't work if it's still doing a task"* (P20). In addition, our participants often considered the INTERVENTIONS inconvenient; for example, when the microphone, camera, and internet are deactivated, there is hardly any way of restarting the robot (P4, P17). Finally, our participants discussed that familiar communication patterns have the big advantage of being immediately understandable (P19), that humanoid patterns are more understandable due to their intuitiveness (P3, P5, P19), and that they would prefer to use patterns they considered fun to use (P7, P10): *"It is fun. Like it's something that is trying to mimic me, but it's not me"* (P7).

4.3.5 Humanoid vs. Non-Humanoid

Our participants discussed that humanoid robots provoke human expectations as their shape makes them appear more capable (P1, P2), which also makes them feel less controllable (P6) and sometimes even evokes feelings of unease (P2, P6, P7, P11, P20): *"I wouldn't want human-like with skin on it or something, because it would be creepy"* (P7). The anthropomorphic appearance also led to people discussing whether the robots would then develop some form of consciousness, evoking feelings of pity (P2, P3, P7): *"Maybe you get emotionally attached in a way that you feel sorry for them when they have to do certain tasks [...] it feels like enslaving"* (P2). Yet, other participants completely disagreed and stated that they would never feel sorry for a machine, regardless of its appearance (P10, P13). Moreover, our participants also discussed that the human-like shape might evoke feelings of trust, which can be unjustified as the robot might collect and share sensitive data (P8). Finally, the participants debated that while some communication patterns are already weird if used by a human, for example, staying in the same room but covering the eyes to signal that one is not watching (P4), this would become even stranger if adopted by a robot (P5): *"If a robot is covering its eyes I would be like: What's wrong with you? Just turn off your camera, dude!"*

4.4 Gesture Set Extraction

To construct the gesture set, we reviewed all individual quotes in the themes INTERVENTIONS and AWARENESS MECHANISMS and merged all quotes that described the same communication pattern. We further categorized the remaining quotations by their tackled functionality, i.e., camera, microphone, or internet connectivity. This process resulted in 86 individual communication patterns, 33 INTERVENTIONS and

53 AWARENESS MECHANISMS. Twenty-eight tackled the camera, 27 the microphone, 21 the internet connectivity, and 10 all functionalities simultaneously. Please refer to Tab. 1 for the complete list of all communication patterns.

5 Study III: Evaluating the Patterns

Via a large-scale online survey, we determined which patterns performed best regarding trust, privacy, understandability, notification quality, and general user preference (**RQ3**). Our ethics committee approved the survey.

5.1 Survey Construction

The survey started with a short introductory text, instructing the participants to immerse themselves in a future situation where they own a domestic robot that supports them with daily chores. The text further stated that the robot uses a communication pattern to show that the user's privacy is protected. After that, every participant saw one of the 86 communication patterns. For INTERVENTIONS, we used the following sentence structure: `The domestic robot does [communication pattern] to physically prevent [capability]`, and for AWARENESS MECHANISMS, we used: `The domestic robot does [communication pattern] to signal that [capability] is deactivated`. Next, we asked them to rate eight statements on a 100-point scale (from strongly disagree to strongly agree). We used VAS without ticks for the same reasons as previously stated [19, 37, 43]. We asked (1) how well our participants felt their privacy was protected, (2) how much they trusted the capability to be actually deactivated, (3) how effective, (4) intrusive, (5) noticeable, (6) understandable, and (7) disturbing they considered the communication pattern and finally, (8) how much the participant would like their domestic robot to use the communication pattern. Additionally, we asked them to put the slider all the way to the right side as an attention check. We used the statements of Rzayev et al. [44] to investigate the notification quality (statements (3) to (7)) in line with [51]. For the full questionnaire, see Sec. A.3.

5.2 Participants

We recruited 1720 participants via Prolific as we wanted to have 20 ratings per communication pattern. We used no reputational filters, and our participants had a mean of 490 ($SD = 534$) approved tasks. We recruited our participants in several batches to (1) replace participants who failed the attention check (see Sec. A.3, question 7) ($N = 2$) and (2) counterbalance the sample in terms of country of birth and gender. The participants were between 18 and 71 ($M = 34.6$, $SD = 9.5$) years old, and 869 identified male, 825 as female, 22 as non-binary, and four did not disclose their gender. 1665

were full-time, and 55 were employed part-time, of whom 107 were also students. Most held an undergraduate degree (659), a graduate degree (585), or a high school diploma (208). Our participants were born in 107 different countries. Most had their origin in the UK (123), Poland (102), Portugal (87), Italy (86), South Africa (84), and Mexico (83). We compensated the 1 min survey with 0.15£.

5.3 Results

We analyzed our data using Python. First, we employed hierarchical clustering to understand the underlying relationships among the communication patterns. This allowed us to build clusters based on linkage criteria and distance thresholds. Thereby, we found three distinct clusters: one consisting of 80 communication patterns, one of five, and one cluster that only contained a single communication pattern. We used principal component analysis (PCA) to reduce the eight measurements (*Privacy*, *Trust*, *Effectiveness*, *Intrusiveness*, *Noticability*, *Understandability*, *Disturbance*, and *Preference*) to two dimensions for easier investigation; see Fig. 2. The PCA visualization shows that the big cluster is separated from the two other clusters. To understand the meaning of our clusters, we utilized parallel coordinates plots where we highlighted the separate clusters. Here, Fig. 2a revealed that the two "outlier" clusters comprise all low-scoring communication patterns. Five of these "outlier" patterns are represented with a similar curve in the parallel coordinates plot, showing that they scored equally low regarding privacy, trust, and user preference. Those patterns were: (1) the robot covering its ears with its hands (CP10), (2) or facing the wall to prevent audio recordings (CP36), (3) the robot deactivating its rotation function to signal that the camera is off (CP38), (4) the robot facing the wall to signal that the microphone is off (CP38), and (5) the robot parking against a pillow to prevent the microphone from recording (CP46), whereby CP36 and CP46 scored lowest regarding trust and privacy. CP45, the robot killing itself to prevent all capabilities, behaved differently than all other patterns and was perceived as, by far, the most disturbing. Yet, it scored well regarding privacy and trust. *We attribute the low scores of these patterns to either their inability to convincingly block a sensor, such as parking against a pillow to interfere with the microphone state, or to the disconnect between the action and targeted capability, such as facing a wall to signal microphone states. Finally, the robot covering its ears might have been perceived as strange or deceptive, and the robot killing itself scored low overall because of its extremely disturbing nature.*

Moreover, we also highlighted the best-scoring patterns in Fig. 2b. Their opposing position with respect to the low-scoring patterns indicates that the PCA can capture the quality of the patterns. Comparing the insights from both plots, we see that while we found some outliers, most patterns were equally well received. Eight out of the ten best scoring pat-

terns are interventions, i.e., actions done by the robot that physically prevent the capability. In detail, the best scoring patterns were: (1) the robot putting a physical cover over its camera (CP51), the robot blocking its own movement (CP6), the robot deactivating its rotation function (CP15), or the robot using a physical switch (CP69) to prevent the camera from recording; the robot removing the microphone's cable (CP56) or detaching the microphone (CP19) to prevent audio recordings; and the robot detaching its memory card (CP1), or going to its docking station (CP39) to prevent all functionalities at once. In contrast, the two best-scoring awareness mechanisms are both human gestures, whereby one was more generally phrased: The robot uses a hand gesture to signal that the microphone is off (CP68), and the other one very concretely: The robot crossing its arm to signal that it is disconnected from the internet (CP14). *In summary, most patterns that scored well across all measurements represented interventions that are familiar from the smart home environment (i.e., a camera shutter or going to the docking station) or represent interventions a human would do but applied to the robot (i.e., removing the cable or memory card, detaching the sensor [23]).*

In Fig. 3, we visualize each pattern's average score for the *Privacy* measurement. Here, we see that the three best-performing patterns are all interventions, meaning they not only signal the sensor state but physically prevent the functionality. In detail, the three best-performing patterns in regards to *Privacy* are (1) the robot putting a physical cover over its camera to prevent it from filming (CP51), (2) the robot detaching its microphone (CP19), and (3) the robot removing the microphone's cable (CP56) to prevent audio recordings. In contrast, the three worst-performing patterns are (1) the robot facing the wall to prevent the camera from filming (CP36), (2) the robot covering its ears with its hands to prevent the microphone from functioning (CP10), and (3) the robot parking against a pillow to prevent audio recordings (CP46). While these patterns are also all interventions, they represent more experimental and unfamiliar patterns. In addition, CP10 has a very large interquartile range (IQR), showing how differently our participants perceived the pattern. Moreover, the rather large IQRs across all communication patterns ($M = 50.2$, $SD = 29.4$) quantify their polarizing nature. *We find that seven of the overall best scoring patterns also scored best regarding their mean privacy rating. This shows, on the one hand, the small differences between the patterns and that many scored almost equally well. On the other hand, this shows a high disparity between the measurements, meaning that while a pattern can be perceived as very privacy-preserving, it might not score as well regarding the other measurements, signifying the importance of choosing the right pattern for a specific goal or situation.*

For an overview of all patterns' means and SDs, see Tab. 1. We created an interactive web app (https://robot-patterns-finder.web.app/) that allows designers and researchers to explore communication patterns based on various requirements.
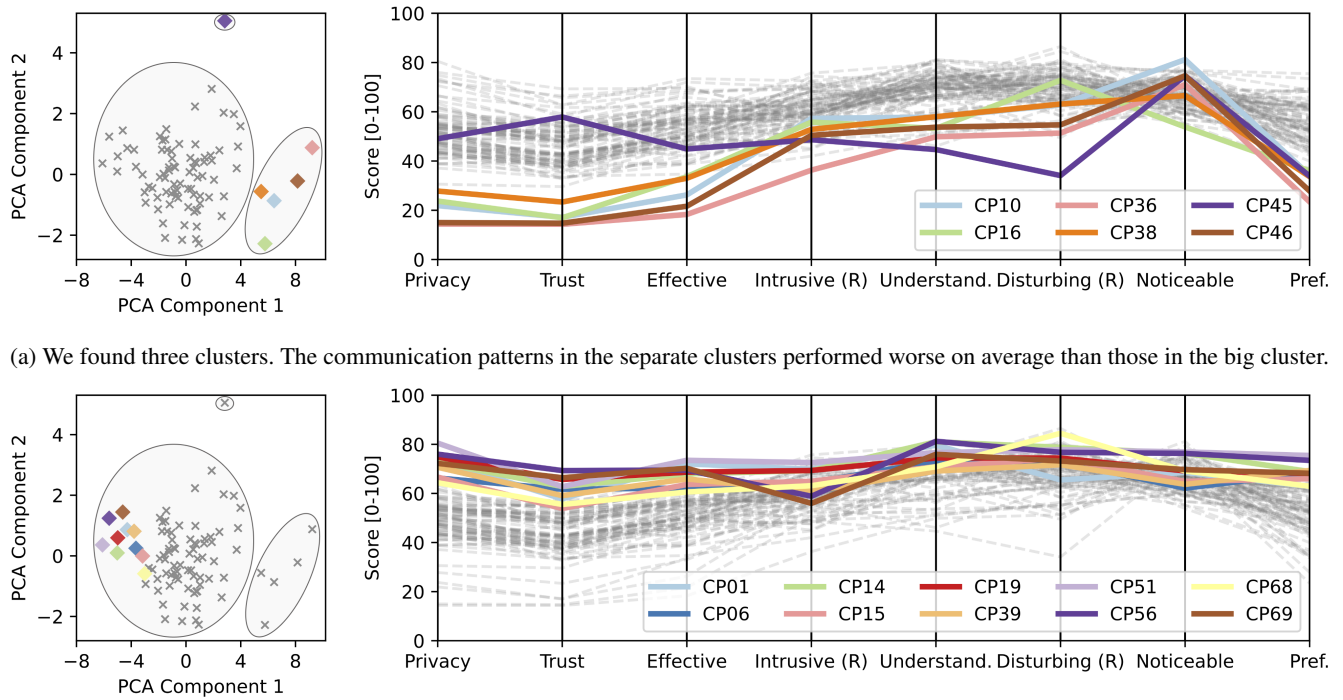
(a) We found three clusters. The communication patterns in the separate clusters performed worse on average than those in the big cluster.

(b) The ten best-performing patterns highlighted.

Figure 2: Insights into the communication patterns. We reversed the two negative items for semantic readability (R).

6 Discussion

We found that domestic robots' increasing locomotion and interaction capabilities lead to heightened privacy concerns (**RQ1**), that their novel interaction and locomotion capabilities enable new ways to indicate or intervene with their sensor states (**RQ2**), and that most communication patterns perform equally well, showing that pattern use depends on the specific requirements of a situation (**RQ3**). In the following, we will discuss and relate our key findings to prior work.

6.1 Interventions for Advanced Capabilities

While prior work warned about the privacy threats rooted in domestic robots' increased mobility and physicality [11, 32], there is no work so far linking privacy concerns directly with those capabilities. Quite the contrary, prior work even found that users are only mildly concerned about their physical privacy when dealing with domestic robots [31]. In contrast to this, we found that participants' privacy concerns increase step-wise with rising interaction and locomotion. Our participants explained their increased concerns with loss of control: While, in the case of stationary robots, they could still restrict what the robot could hear and see by placing it in specific areas, robotic systems with various locomotion and interaction capabilities can search through private documents or even unlock doors, leaving virtually no space for privacy.

This was also picked up in the focus groups, where our participants agreed that advanced robot capabilities require stronger communication patterns. Here, our participants suggested awareness mechanisms most frequently for stationary robots with limited interaction capabilities, such as simple light indications or audio feedback. At the same time, they wished for the highest level of privacy when dealing with robots with advanced capabilities. Here, our participants' suggestions most often included intervention mechanisms, but even those were sometimes not perceived as secure enough. As a result, their suggestions also included ways to stop the robot from recovering its functional state, such as adding physical locks to prevent it from leaving a physical enclosure or moving detached sensors and cables out of the robot's reach. ***Key Finding 1: Advanced Capabilities Require Strong Interventions. The more capable a domestic robot is, the more it threatens users' privacy, evoking the desire for mechanisms that provide the highest levels of certainty and trust.***

6.2 Familiarity for Understanding and Trust

Our results show that most of the well-scoring patterns either represent familiar interventions, such as physical covers or entering the docking station, or interventions usually employed by humans to mitigate their concerns, such as unplugging cables [23]. We attribute the high scores of these patterns to their tangibility and familiarity, making it easy for users to
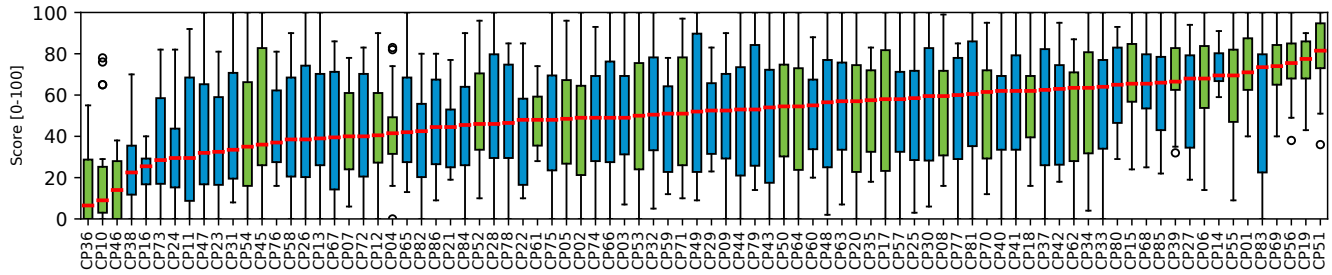
Figure 3: Mean ratings for the PRIVACY statement. Interventions are green, and awareness mechanisms are blue.

understand how they work. In fact, prior work emphasized the value of employing tangible mechanisms for higher trust and understandability [3], which ultimately contributes to inclusive privacy [52]. Yet, this relationship between familiarity and trust also works the other way around; some patterns scored low as users felt they might not be effective. For example, preventing audio recordings by facing the wall or parking against a pillow. We attribute the low scores to users being aware of the high sensitivity of current audio sensors that can capture noises even when obstructed. Yet, the advantage of familiarity is not only true for tangible mechanisms. Also, human hand gestures scored well in our third study. This can be explained by discussions from our focus group, where participants praised these gestures for being understandable and intuitive. ***Key Finding 2: Familiarity with a pattern fosters understandability, trust, and general user preference.*** *Such familiarity can stem from smart devices already having similar mechanisms integrated or from applying knowledge and actions from daily life to the novel robotics space.*

6.3 Humanoid Robots and Metaphors

In contrast to our participants' general preference for humanoid hand gestures, other patterns leveraging human metaphors performed badly, such as the robot covering its ears to prevent audio recordings. Our focus groups can explain this. Here, participants discussed that they would find it even weird for humans to cover their ears to signal that they are not listening instead of simply leaving the room. Hence, a robot replicating such behavior would be even more strange. Another reason might be the difference between awareness mechanisms and interventions. While signifying the sensor state using hand gestures might be well understandable and, thus, well received, covering the ears as an intervention mechanism might provoke distrust; users might be skeptical that the gesture prevents the recording capability, especially as the robot's microphones are not necessarily placed in the ear.

Our focus group participants also discussed that the robot's shape influences their general perception; they agreed that a humanoid shape makes a robot seem more capable. At the same time, however, they also discussed that a humanoid

form makes a robot seem less controllable. Some participants even considered a too-humanoid appearance creepy, linking to the well-recognized uncanny valley effect [45], and discussed that their shape might evoke undesired feelings, such as feeling pity for the robot when it has to complete undesired tasks. In this regard, prior work suggested exploring the value of "honest anthropomorphism," meaning using anthropomorphic features to notify the users of what a robot is actually doing [24]. Our results show that while anthropomorphic patterns can help foster understandability and trust, they are sometimes perceived as creepy or weird. Moreover, we found them to be more suitable for awareness mechanisms than for interventions. ***Key finding 3: While humanoid shapes and behaviors foster understandability through intuitiveness and familiarity, they can also evoke feelings of unease and even creepiness.*** *Hence, we suggest employing anthropomorphism carefully and align it with the specific situation.*

6.4 Choosing the Right Pattern

In summary, many factors must be considered when choosing the optimal communication pattern. As discussed previously, the more capable and intrusive a robot is, the stronger the employed interventions should be. Similarly, Windl et al. [53] suggest that preventing a situation from being privacy violating should be preferred (i.e., through interventions) in contrast to using notices (i.e., awareness mechanisms) whenever possible. Yet, they also discuss that the right mechanism strongly depends on the constraints of a situation. This is especially true in the case of domestic robots, as it is often not as easy as unplugging the robot or sending it away. In contrast, the robot most often needs its full capabilities to fulfill the tasks it was purchased for in the first place. Hence, which communication pattern to employ also depends on the robots' task and whether it is currently actively working or not. That means that, even though interventions provide higher levels of trust and certainty, sometimes awareness mechanisms might be the better option. Moreover, while familiar patterns are often perceived as very understandable and trustworthy, and using humanoid metaphors should certainly be considered familiar, their usage must still be carefully considered as they walk a

fine line between being intuitive and creepy.

The varying individual ratings also reflect this discrepancy and polarizing nature of some communication patterns. While the measurements for privacy, trust, and overall user preference seem to mostly correlate (see Fig. 2), the other measurements do not seem to follow a similar pattern: While a communication pattern might convey high levels of privacy and trust, it might also be perceived as disturbing or barely noticeable. In addition, the high variance speaks for a generally highly subjective perception of some patterns. As we recognized this discrepancy between the different measurements and that the importance of individual measurements depends on the characteristics of a situation, we created an interactive web application that allows researchers and developers to filter our extensive pattern set depending on their needs. ***Key Finding 4: Choosing the right communication pattern does not follow a simple one-size-fits-all approach; in contrast, which communication is best depends on the specific requirements of a situation.***

6.5   Limitations and Future Work

We used an online survey to understand users' privacy concerns towards domestic robots with increasing capabilities. While online surveys are an established method to elicit privacy concerns [31, 50], and sometimes the only viable option when investigating future scenarios, they might suffer from biases caused by participants having to immerse themselves in the described future or participants indicating answers that might not reflect their actual behavior [25]. In real life, participants might be more considerate of the convenience provided by the robot, making them willing to trade some of their privacy for an increased quality of life [17]. Moreover, the generally high privacy concerns might also be attributed to participants' low familiarity with such robots. Indeed, prior work already showed that higher familiarity is linked to decreased privacy concerns [6, 50]. Consequently, it will be interesting to repeat our survey in the future to see how concerns shift as users become familiar with domestic robots.

For this investigation, we did not consider the technical feasibility or how easy the gestures are to implement; we only focused on the users' perspective and which patterns provoke the highest levels of trust. Yet, in practice, technical feasibility is an important factor to consider when deciding which communication pattern to adopt. Hence, we recommend that future work employs a more technical focus and discusses the feasibility of our retrieved patterns from this perspective.

We limited our elicitation of communication patterns to the three most privacy-concerning capabilities. We argue that limiting our investigation was important to be able to conduct the studies. Moreover, offering interventions and communicating the state of the most privacy-relevant capabilities is an approach frequently followed by manufacturers – many smart device manufacturers only provide mechanisms to physically block the cameras or integrate hardware buttons to deactivate the microphone. Yet, in reality, smart home appliances, and especially future domestic robots, will have way more privacy-relevant sensors, and which sensors are perceived as privacy-relevant might differ by user. Thus, it will be interesting to investigate which of our patterns apply to a broader range of sensors and where we need new mechanisms. Moreover, as previously discussed, concerns go beyond the pure collection of data as outlined in Solove's [47] taxonomy of privacy harms. Hence, future investigations are needed following this taxonomy as prior research already did for less capable smart assistants, c.f. [1, 2].

We showed the focus group participants examples, i.e., a mute button and a physical camera shutter, to clarify what we mean by communication patterns. While our results show that our participants came up with a wide range of diverse patterns, we still want to acknowledge that these examples might have introduced unintentional biases as we can not exclude that other examples, such as LEDs [39] or dialogues with the user [57], might have led to different or more diverse communication patterns.

Lastly, we used an online survey to describe the communication patterns in Study III. While we are certain that this is a good approach to get a first impression of the feasibility of the gestures, and online surveys are also a typical method used to gather human's perception towards robots [29], how the patterns are actually perceived in real life might be different. Hence, it would be desirable to test a selection of the patterns using prototypes, for example, in a lab study setting.

7   Conclusion

We conducted three studies: An online survey (N=90), a focus group study (N=22), and a final large-scale online survey (N=1720) to understand users' privacy concerns towards future domestic robots and develop communication patterns to intervene with and signify their sensor state. Through this, we found that (1) the more interaction and movement capabilities a domestic robot has, the more concerns it evokes; (2) these novel capabilities also enable completely new communication patterns; and (3) most of these diverse patterns score equally well across all measurements, meaning that pattern use depends on the situation. To help researchers and developers navigate our extensive set of communication patterns along the mentioned characteristics, we developed an interactive web app. Finally, we discuss our key insights for choosing the right communication pattern: (1) selecting the mechanism based on the robot's capabilities, (2) choosing familiar patterns whenever possible to foster understandability and trust, and (3) being wary of the potential pitfalls when using humanoid metaphors.

References

[1] Noura Abdi, Kopo M. Ramokapane, and Jose M. Such. More than Smart Speakers: Security and Privacy Perceptions of Smart Home Personal Assistants. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*. USENIX Association, 2019. URL https://www.usenix.org/conference/soups2019/presentation/abdi.

[2] Noura Abdi, Xiao Zhan, Kopo M. Ramokapane, and Jose Such. Privacy Norms for Smart Home Personal Assistants. In *Proc. of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21. ACM, 2021. doi: 10.1145/3411764.3445122.

[3] Imtiaz Ahmad, Rosta Farzan, Apu Kapadia, and Adam J. Lee. Tangible Privacy: Towards User-Centric Sensor Designs for Bystander Privacy. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2), 2020. doi: 10.1145/3415187.

[4] Noah Apthorpe, Dillon Reisman, and Nick Feamster. A smart home is no castle: Privacy vulnerabilities of encrypted iot traffic. *Workshop on Data and Algorithmic Transparency*, 2016.

[5] Noah Apthorpe, Dillon Reisman, Srikanth Sundaresan, Arvind Narayanan, and Nick Feamster. Spying on the Smart Home: Privacy Attacks and Defenses on Encrypted IoT Traffic. *arXiv preprint arXiv:1708.05044*, 2017. doi: 10.48550/arXiv.1708.05044.

[6] Noah Apthorpe, Yan Shvartzshnaider, Arunesh Mathur, Dillon Reisman, and Nick Feamster. Discovering Smart Home Internet of Things Privacy Norms Using Contextual Integrity. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(2), 2018. doi: 10.1145/3214262.

[7] Abdullahi Arabo, Ian Brown, and Fadi El-Moussa. Privacy in the Age of Mobility and Smart Devices in Smart Homes. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, 2012. doi: 10.1109/SocialCom-PASSAT.2012.108.

[8] Ann Blandford, Dominic Furniss, and Stephann Makri. *Qualitative HCI Research: Going Behind the Scenes*. Synthesis Lectures on Human-Centered Informatics. Springer Cham, 2016. doi: 10.2200/S00706ED1V01Y201602HCI034.

[9] Joseph Bugeja, Andreas Jacobsson, and Paul Davidsson. On privacy and security challenges in smart connected homes. In *2016 European Intelligence and Security Informatics Conference*, EISIC, 16. IEEE, 2016. doi: 10.1109/EISIC.2016.044.

[10] M. Ryan Calo. Peeping hals. *Artificial Intelligence*, 175(5), 2011. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2010.11.025.

[11] Ryan Calo. Robotics and the lessons of cyberlaw. *California Law Review*, 103(3), 2015. ISSN 00081221. URL http://www.jstor.org/stable/24758483.

[12] George Chalhoub, Martin J Kraemer, Norbert Nthala, and Ivan Flechais. "It did not give me an option to decline": A Longitudinal Analysis of the User Experience of Security and Privacy in Smart Home Products. In *Proc. of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21. ACM, 2021. doi: 10.1145/3411764.3445691.

[13] Eun Kyoung Choe, Sunny Consolvo, Jaeyeon Jung, Beverly Harrison, Shwetak N. Patel, and Julie A. Kientz. Investigating Receptiveness to Sensing and Inference in the Home Using Sensor Proxies. In *Proc. of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12. ACM, 2012. doi: 10.1145/2370216.2370226.

[14] Sarah Delgado Rodriguez, Sarah Prange, and Florian Alt. Take Your Security and Privacy Into Your Own Hands! Why Security and Privacy Assistants Should be Tangible. In Carolin Wienrich, Philipp Wintersberger, and Benjamin Weyers, editors, *Mensch und Computer - Workshopband*. Gesellschaft für Informatik e.V., 2021. doi: https://doi.org/10.18420/muc2021-mci-ws09-393.

[15] Tamara Denning, Cynthia Matuszek, Karl Koscher, Joshua R. Smith, and Tadayoshi Kohno. A spotlight on security and privacy risks with future household robots: attacks and lessons. In *Proc. of the 11th International Conference on Ubiquitous Computing*, UbiComp '09. ACM, 2009. doi: 10.1145/1620545.1620564.

[16] Robert F DeVellis and Carolyn T Thorpe. *Scale development: Theory and applications*. SAGE, 2021.

[17] Tamara Dinev and Paul Hart. An extended privacy calculus model for e-commerce transactions. *Information systems research*, 17(1), 2006.

[18] Thomas Franke, Christiane Attig, and Daniel Wessel. A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale. *International Journal of Human–Computer Interaction*, 35(6), 2019. doi: 10.1080/10447318.2018.1456150.

[19] Frederik Funke and Ulf-Dietrich Reips. Why Semantic Differentials in Web-Based Research Should Be Made from Visual Analogue Scales and Not from 5-Point Scales. *Field Methods*, 24(3), 2012. doi: 10.1177/1525822X12444061.

---

[20] Nina Gerber, Benjamin Reinheimer, and Melanie Volkamer. Home Sweet Home? Investigating Users' Awareness of Smart Home Privacy Threats. In *Proc. of An Interactive Workshop on the Human aspects of Smarthome Security and Privacy (WSSP)*. USENIX, 2018. doi: 10.5445/IR/1000083578.

[21] Gunnar Harboe and Elaine M. Huang. Real-World Affinity Diagramming Practices: Bridging the Paper-Digital Gap. In *Proc. 33rd Annual ACM Conf. Human Factors in Computing Systems*. ACM, 2015. doi: 10.1145/2702123.2702561.

[22] Roberto Hoyle, Luke Stark, Qatrunnada Ismail, David Crandall, Apu Kapadia, and Denise Anthony. Privacy Norms and Preferences for Photos Posted Online. *ACM Trans. Comput.-Hum. Interact.*, 27(4), 2020. doi: 10.1145/3380960.

[23] Haojian Jin, Boyuan Guo, Rituparna Roychoudhury, Yaxing Yao, Swarun Kumar, Yuvraj Agarwal, and Jason I. Hong. Exploring the Needs of Users for Supporting Privacy-Protective Behaviors in Smart Homes. In *Proc. of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22. ACM, 2022. doi: 10.1145/3491102.3517602.

[24] Margot E Kaminski, Matthew Rueben, William D Smart, and Cindy M Grimm. Averting robot eyes. *Md. L. Rev.*, 76, 2016.

[25] Spyros Kokolakis. Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon. *Computers & Security*, 64, 2017. doi: 10.1016/j.cose.2015.07.002.

[26] Evan Lafontaine, Aafaq Sabir, and Anupam Das. Understanding People's Attitude and Concerns towards Adopting IoT Devices. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI'21. ACM, 2021. doi: 10.1145/3411763.3451633.

[27] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. Alexa, Are You Listening? Privacy Perceptions, Concerns and Privacy-Seeking Behaviors with Smart Speakers. *Proc. ACM Hum.-Comput. Interact.*, 2 (CSCW), 2018. doi: 10.1145/3274371.

[28] Min Kyung Lee, Karen P. Tang, Jodi Forlizzi, and Sara Kiesler. Understanding Users' Perception of Privacy in Human-Robot Interaction. In *Proc. of the 6th International Conference on Human-Robot Interaction*, HRI '11. ACM, 2011. doi: 10.1145/1957656.1957721.

[29] Jan Leusmann, Carl Oechsner, Johanna Prinz, Robin Welsch, and Sven Mayer. A Database for Kitchen Objects: Investigating Danger Perception in the Context of Human-Robot Interaction. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI EA '23. ACM, 2023. doi: 10.1145/3544549.3585884.

[30] Huichen Lin and Neil W. Bergmann. IoT Privacy and Security Challenges for Smart Home Environments. *Information*, 7(3), 2016. doi: 10.3390/info7030044.

[31] Christoph Lutz and Aurelia Tamó-Larrieux. The robot privacy paradox: Understanding how privacy concerns shape intentions to use social robots. *Human-Machine Communication*, 1, 2020. doi: 10.30658/hmc.1.6.

[32] Christoph Lutz, Maren Schöttler, and Christian Pieter Hoffmann. The privacy implications of social robots: Scoping review and expert interviews. *Mobile Media & Communication*, 7(3), 2019. doi: 10.1177/2050157919843961.

[33] Naresh K. Malhotra, Sung S. Kim, and James Agarwal. Internet Users' Information Privacy Concerns (IUIPC): The Construct, the Scale, and a Causal Model. *Information Systems Research*, 15(4), 2004. doi: 10.1287/isre.1040.0032.

[34] Nathan Malkin, Julia Bernd, Maritza Johnson, and Serge Egelman. "What Can't Data Be Used For?" Privacy Expectations about Smart TVs in the US. In *Proc. of the 3rd European Workshop on Usable Security (EuroUSEC)*, 2018. doi: 10.14722/eurousec.2018.23016.

[35] Nathan Malkin, Joe Deatrick, Allen Tong, Primal Wijesekera, Serge Egelman, and David Wagner. Privacy attitudes of smart speaker users. *Proc. on Privacy Enhancing Tech.*, 2019. doi: 10.2478/popets-2019-0068.

[36] Shrirang Mare, Franziska Roesner, and Tadayoshi Kohno. Smart Devices in Airbnbs: Considering Privacy and Security for both Guests and Hosts. *Proc. on Privacy Enhancing Technologies*, 2020(2), 2020. doi: 10.2478/popets-2020-0035.

[37] Justin Matejka, Michael Glueck, Tovi Grossman, and George Fitzmaurice. The Effect of Visual Appearance on the Performance of Continuous Sliders and Visual Analogue Scales. In *Proc. of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16. ACM, 2016. doi: 10.1145/2858036.2858063.

[38] Vikram Mehta, Daniel Gooch, Arosha Bandara, Blaine Price, and Bashar Nuseibeh. Privacy Care: A Tangible Interaction Framework for Privacy Management. *Trans. Internet Technol.*, 21(1), 2021. doi: 10.1145/3430506.

[39] Abraham Mhaidli, Manikandan Kandadai Venkatesh, Yixin Zou, and Florian Schaub. Listen only when spoken to: Interpersonal communication cues as smart speaker

privacy controls. *Proc. on Privacy Enhancing Technologies*, 2020. doi: 10.2478/popets-2020-0026.

[40] Andrés Molina-Markham, Prashant Shenoy, Kevin Fu, Emmanuel Cecchet, and David Irwin. Private Memoirs of a Smart Meter. In *Proc. of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building*, BuildSys '10. ACM, 2010. doi: 10.1145/1878431.1878446.

[41] Simon Moncrieff, Svetha Venkatesh, and Geoff West. Dynamic Privacy in a Smart House Environment. In *2007 IEEE International Conference on Multimedia and Expo*, 2007. doi: 10.1109/ICME.2007.4285080.

[42] Johannes Obermaier and Martin Hutle. Analyzing the Security and Privacy of Cloud-Based Video Surveillance Systems. In *Proc. of the 2nd ACM International Workshop on IoT Privacy, Trust, and Security*, IoTPTS '16. ACM, 2016. doi: 10.1145/2899007.2899008.

[43] Ulf-Dietrich Reips and Frederik Funke. Interval-level measurement with visual analogue scales in Internet-based research: VAS Generator. *Behavior Research Methods*, 40(3), 2008. doi: 10.3758/BRM.40.3.699.

[44] Rufat Rzayev, Sven Mayer, Christian Krauter, and Niels Henze. Notification in VR: The Effect of Notification Placement, Task and Environment. In *Proc. of the Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '19. ACM, 2019. doi: 10.1145/3311350.3347190.

[45] Jun'ichiro Seyama and Ruth S. Nagayama. The Uncanny Valley: Effect of Realism on the Impression of Artificial Human Faces. *Presence*, 16(4), 2007. doi: 10.1162/pres.16.4.337.

[46] Noel Sharkey and Amanda Sharkey. The eldercare factory. *Gerontology*, 58(3), 2012. doi: 10.1159/000329483.

[47] Daniel J Solove. A taxonomy of privacy. *University of Pennsylvania Law Review*, 154, 2005. doi: 10.2307/40041279.

[48] Meg Tonkin, Jonathan Vitale, Suman Ojha, Jesse Clark, Sammy Pfeiffer, William Judge, Xun Wang, and Mary-Anne Williams. Embodiment, Privacy and Social Robots: May I Remember You? In *Social Robotics: 9th International Conference*, ICSR 2017. Springer, 2017. doi: 10.1007/978-3-319-70022-9_50.

[49] Steeven Villa, Jasmin Niess, Takuro Nakao, Jonathan Lazar, Albrecht Schmidt, and Tonja-Katrin Machulla. Understanding Perception of Human Augmentation: A Mixed-Method Study. In *Proc. of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23. ACM, 2023. doi: 10.1145/3544548.3581485.

[50] Maximiliane Windl and Sven Mayer. The Skewed Privacy Concerns of Bystanders in Smart Environments. *Proc. ACM Hum.-Comput. Interact.*, 6(MHCI), 2022. doi: 10.1145/3546719.

[51] Maximiliane Windl, Anna Scheidle, Ceenu George, and Sven Mayer. Investigating security indicators for hyperlinking within the metaverse. In *Nineteenth Symposium on Usable Privacy and Security (SOUPS 2023)*. USENIX Association, 2023. URL https://www.usenix.org/conference/soups2023/presentation/windl.

[52] Maximiliane Windl, Albrecht Schmidt, and Sebastian S. Feger. Investigating Tangible Privacy-Preserving Mechanisms for Future Smart Homes. In *Proc. of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23. ACM, 2023. doi: 10.1145/3544548.3581167.

[53] Maximiliane Windl, Verena Winterhalter, Albrecht Schmidt, and Sven Mayer. Understanding and Mitigating Technology-Facilitated Privacy Violations in the Physical World. In *Proc. of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23. ACM, 2023. doi: 10.1145/3544548.3580909.

[54] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only Anova Procedures. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11. ACM, 2011. doi: 10.1145/1978942.1978963.

[55] Yaxing Yao, Justin Reed Basdeo, Smirity Kaushik, and Yang Wang. Defending My Castle: A Co-Design Study of Privacy Mechanisms for Smart Homes. In *Proc. of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19. ACM, 2019. doi: 10.1145/3290605.3300428.

[56] Yaxing Yao, Justin Reed Basdeo, Oriana Rosata Mcdonough, and Yang Wang. Privacy Perceptions and Designs of Bystanders in Smart Homes. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), 2019. doi: 10.1145/3359161.

[57] Nicole Zhan, Stefan Sarkadi, and Jose Such. Privacy-enhanced Personal Assistants based on Dialogues and Case Similarity. In *European Conference on Artificial Intelligence*. IOS Press, 2023.

[58] Serena Zheng, Noah Apthorpe, Marshini Chetty, and Nick Feamster. User Perceptions of Smart Home IoT Privacy. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW), 2018. doi: 10.1145/3274469.

A Appendix

A.1 Survey on Privacy Concerns

1. Demographics

2. IUIPC

3. ATI

4. *[Main part of the survey, repeated 12 times for every locomotion + interaction combination in random order.]* Imagine the following scenario – You own a smart assistant that you are using in your home. It has the following capabilities: *[Capability Description.]* Please indicate to which extent you agree with the following statement:

   (a) I am strongly concerned about my privacy due to the presence of the smart assistant. (Slider)

   (b) Please explain your reasoning for the above answer. (Free text)

   (c) Please move the slider all the way to the [left/right]. (Attention Check)

5. If you have any further feedback regarding this situation, you can let us know here. (Free text)

A.2 Communication Patterns

Table 1: All communication patterns that resulted from the focus groups, whether they are an AWARENESS MECHANISMS or an INTERVENTION and which sensor they tackle. In addition, the table contains the means (M) and standard deviations (SD) for all measurements that resulted from Study III: Privacy (Pri.), Trust (T), Effectiveness (E), Intrusiveness (I), Noticability (N), Understandability (U), Disturbance (D), Preference (Pref.)

Q	Communication Pattern	Pri. M	Pri. SD	T M	T SD	E M	E SD	I M	I SD	N M	N SD	U M	U SD	D M	D SD	Pref. M	Pref. SD
CP01	The domestic robot detaches its memory card to physically prevent the camera, microphone, and internet connection from functioning.	73.6	18.2	57.8	26.6	72.0	20.2	30.2	19.0	68.6	23.4	79.7	19.9	34.5	25.1	62.8	26.6
CP02	The domestic robot moves out of the WiFi range to physically prevent the internet connection.	43.8	31.6	33.8	24.2	34.7	26.1	58.0	19.9	59.9	20.4	51.3	27.0	44.6	27.9	43.4	29.8
CP03	The domestic robot retracts its camera to signal that the camera is off.	49.6	27.2	48.2	29.9	56.2	25.4	39.4	26.8	67.0	22.6	66.6	23.6	27.1	23.0	62.2	22.8
CP04	The domestic robot shows you the empty connection plug to physically prevent the internet connection.	42.9	21.8	37.4	25.7	46.4	25.3	50.9	27.3	65.6	21.2	59.4	24.8	37.6	30.7	45.3	25.9
CP05	The domestic robot turns its screen off to physically prevent the camera, microphone, and internet connection from functioning.	47.5	28.3	41.8	30.5	46.0	26.4	25.5	26.4	67.2	21.3	72.2	17.6	21.8	25.4	58.5	23.4
CP06	The domestic robot blocks its own movement to physically prevent the camera from recording.	66.3	25.1	61.7	25.3	62.2	25.0	35.6	26.7	62.1	21.4	73.1	18.0	27.0	20.6	68.5	24.7
CP07	The domestic robot blocks its own movement to physically prevent the internet connection.	41.8	23.8	32.8	23.4	42.2	18.5	44.5	24.3	55.2	21.8	46.0	25.5	37.3	19.5	37.0	23.9
CP08	The domestic robot blocks its own movement to physically prevent the microphone from recording.	55.2	26.1	47.6	28.4	54.4	20.4	31.8	26.6	68.2	14.9	64.2	23.9	25.0	25.6	53.4	27.6
CP09	The domestic robot changes its posture to signal that the camera, microphone, and internet connection are deactivated.	49.2	27.4	43.2	30.5	50.4	24.8	36.3	25.4	61.2	22.4	59.1	25.9	33.6	22.7	48.6	26.7
CP10	The domestic robot covers its ears with its hands to physically prevent the microphone from recording.	21.9	26.9	17.2	23.3	26.2	30.1	42.7	29.0	81.2	18.6	57.8	31.8	37.2	31.3	34.2	28.7
CP11	The domestic robot covers its ears with its hands to signal that the microphone is off.	39.0	32.9	39.4	36.0	38.8	31.8	42.2	34.9	75.2	20.3	76.0	23.2	35.7	35.8	47.9	35.1
CP12	The domestic robot covers its eyes with its hands to physically prevent the camera from recording.	43.6	26.9	38.0	31.0	38.7	26.2	55.2	19.6	73.2	18.5	67.2	21.2	46.9	26.8	34.4	26.5
CP13	The domestic robot covers its eyes with its hands to signal that the camera is off.	46.9	28.9	32.9	19.6	46.9	22.2	34.4	30.7	69.7	28.2	73.8	27.6	23.8	25.4	52.0	26.7
CP14	The domestic robot crosses its arms to signal that it is disconnected from the internet.	72.7	9.3	62.8	21.9	68.0	17.3	30.7	21.8	76.4	14.4	81.0	15.0	21.2	21.4	68.9	20.5
CP15	The domestic robot deactivates its rotation function to physically prevent the camera from recording.	66.6	23.0	53.8	28.4	63.6	26.4	35.0	29.5	64.9	25.6	70.9	22.7	25.7	22.0	66.0	26.7
CP16	The domestic robot deactivates its rotation function to signal that the camera is off.	23.8	11.8	17.0	14.1	33.6	21.1	44.2	24.2	53.8	23.9	53.2	26.3	27.2	27.2	36.0	16.3
CP17	The domestic robot detaches its WiFi module to physically prevent the internet connection.	53.4	33.1	53.0	36.0	56.8	29.4	43.6	25.8	64.4	22.6	73.9	18.8	28.0	28.0	57.6	28.1
CP18	The domestic robot detaches its camera to physically prevent the camera from recording.	57.8	23.9	53.8	22.1	60.7	24.3	39.6	21.2	73.6	16.7	66.6	24.1	34.4	22.2	57.7	20.0
CP19	The domestic robot detaches its microphone to physically prevent the microphone from recording.	74.8	13.2	65.7	24.5	68.6	16.7	30.6	22.0	69.2	23.9	74.4	17.4	25.6	20.0	68.2	21.1
CP20	The domestic robot detaches its power source to physically prevent the camera, microphone, and internet connection from functioning.	52.7	30.5	39.6	34.6	47.0	30.8	39.4	29.5	62.8	27.1	72.0	27.5	28.6	29.9	48.3	33.9
CP21	The domestic robot displays a human gesture on its screen to signal that it is disconnected from the internet.	43.1	19.4	39.0	27.0	50.1	17.4	34.6	18.0	56.7	20.8	71.9	25.4	33.2	26.8	52.6	19.3
CP22	The domestic robot displays a human gesture on its screen to signal that the camera is off.	42.7	25.1	37.0	25.2	51.8	28.9	31.2	24.4	66.8	19.7	65.8	18.8	30.9	23.2	54.8	28.3
CP23	The domestic robot displays a human gesture on its screen to signal that the microphone is off.	38.9	25.1	40.6	29.2	45.6	24.4	43.2	24.2	61.1	15.3	66.7	15.7	29.6	27.7	49.9	25.4
CP24	The domestic robot displays a humanoid face that shuts its eyes on its screen to signal that the camera is off.	30.6	23.1	29.6	20.9	44.8	24.4	40.6	25.8	74.3	16.3	77.2	20.1	31.0	28.6	47.8	27.2
CP25	The domestic robot displays a symbol on its screen to signal that it is disconnected from the internet.	51.2	28.5	52.0	34.9	63.3	27.4	43.0	28.0	70.4	19.9	78.2	22.7	26.0	21.7	63.0	27.7
CP26	The domestic robot displays a symbol on its screen to signal that the camera is off.	47.4	32.6	39.8	32.9	53.2	31.4	37.6	27.4	71.0	19.9	78.4	20.2	33.0	26.7	53.2	30.7
CP27	The domestic robot displays a symbol on its screen to signal that the microphone is off.	60.6	25.3	54.0	33.8	57.2	25.6	35.3	21.4	64.7	20.6	74.0	22.5	25.8	24.5	69.6	26.0
CP28	The domestic robot displays text to signal that it is disconnected from the internet.	52.1	32.2	52.4	30.4	52.9	36.3	37.3	27.8	56.7	22.6	65.6	26.3	26.0	20.6	53.6	31.0
CP29	The domestic robot displays text to signal that the camera is off.	51.0	19.4	41.4	22.2	51.4	18.2	40.8	22.1	66.2	16.4	69.8	18.9	34.8	21.2	54.6	18.1
CP30	The domestic robot displays text to signal that the microphone is off.	55.2	29.2	44.9	31.3	62.6	22.6	32.2	27.5	67.2	20.8	80.8	14.3	25.0	24.0	61.6	27.1
CP31	The domestic robot displays the camera state on its screen to signal that the camera is off.	44.4	31.1	42.0	32.2	57.5	26.5	41.7	28.0	64.8	22.4	80.9	19.2	41.6	33.3	44.1	31.5
CP32	The domestic robot displays the connectivity state on the screen to signal that it is disconnected from the internet.	54.8	27.2	47.8	25.4	55.6	24.5	40.9	26.8	62.7	20.2	71.0	18.5	24.6	22.9	64.8	24.2
CP33	The domestic robot displays the microphone state on its screen to signal that the microphone is off.	56.2	28.3	54.3	33.7	54.7	28.4	28.5	22.7	64.8	18.8	75.2	14.5	27.7	23.2	62.4	26.0
CP34	The domestic robot enters physical confinement to physically prevent the camera, microphone, and internet connection from functioning.	58.7	30.3	57.0	30.2	61.5	24.2	40.8	29.5	74.4	21.6	66.6	24.0	30.0	27.1	54.5	27.0
CP35	The domestic robot faces the wall to physically prevent the camera from recording.	51.8	22.7	36.0	24.1	51.4	24.3	41.5	25.4	75.4	19.8	70.2	24.9	36.8	23.7	52.7	26.1
CP36	The domestic robot faces the wall to physically prevent the microphone from recording.	14.4	17.3	14.3	15.6	18.3	16.8	63.7	29.5	71.0	25.3	49.8	30.1	48.6	31.2	23.4	27.3
CP37	The domestic robot faces the wall to signal that the camera is off.	57.8	33.4	45.8	33.5	59.0	35.5	24.3	20.6	76.8	18.9	79.4	15.9	37.0	34.9	54.8	31.1
CP38	The domestic robot faces the wall to signal that the microphone is off.	27.8	22.3	23.4	24.7	33.0	27.1	47.2	24.0	66.6	22.0	58.0	26.9	36.8	27.5	33.6	17.0
CP39	The domestic robot goes to its docking station to physically prevent the camera, microphone, and internet connection from functioning.	70.6	19.3	59.1	25.7	65.9	18.4	39.2	28.3	63.7	26.8	69.1	24.1	28.3	26.0	69.4	20.4
CP40	The domestic robot has a fake antenna attached that illuminates to signal that it is disconnected from the internet.	54.0	26.4	51.8	21.8	56.4	25.4	51.1	22.4	66.7	19.8	69.0	22.8	28.6	19.6	61.8	17.8

Continued on next page

Table 1 – continued from previous page

Q	Communication Pattern	Pri.		T		E		I		N		U		D		Pref.	
		M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
CP41	The domestic robot has a light band attached to signal that it is disconnected from the internet.	58.2	30.6	55.5	33.4	67.2	27.1	28.7	24.8	59.0	24.8	67.2	28.2	22.4	21.7	64.2	24.0
CP42	The domestic robot has a light band attached to signal that the camera is off.	54.6	26.1	55.8	31.4	62.0	20.0	36.8	26.3	67.4	20.5	72.0	18.7	22.6	25.7	65.2	21.5
CP43	The domestic robot has a light band attached to signal that the microphone is off.	47.8	32.1	39.2	30.3	44.1	27.0	43.8	30.3	75.8	15.9	73.8	24.3	20.3	20.4	61.3	25.8
CP44	The domestic robot imitates the human "shh" gesture/puts its finger in front of his mouth to signal that the microphone is off.	46.6	30.5	37.6	31.7	48.4	30.2	38.6	24.4	61.6	23.7	71.2	19.9	29.2	28.7	48.0	25.8
CP45	The domestic robot kills itself to physically prevent the camera, microphone, and internet connection from functioning.	49.0	34.6	57.9	25.6	44.9	33.3	51.4	32.2	74.6	19.3	44.6	32.4	65.9	38.5	33.9	36.5
CP46	The domestic robot parks itself against a pillow to physically prevent the microphone from recording.	15.0	13.6	14.8	16.6	21.6	18.3	49.6	29.4	74.6	22.4	53.8	29.7	45.3	29.5	27.8	25.4
CP47	The domestic robot plays distinct audio feedback to signal that it is disconnected from the internet.	41.2	31.4	41.2	31.9	50.6	30.6	55.0	29.6	72.2	21.2	70.2	25.5	31.6	29.1	51.2	29.4
CP48	The domestic robot plays distinct audio feedback to signal that the camera is off.	53.8	30.6	51.7	34.3	59.8	29.2	41.4	28.0	73.1	17.6	75.1	20.9	30.1	21.0	50.1	28.8
CP49	The domestic robot plays distinct audio feedback to signal that the microphone is off.	55.6	34.8	49.8	33.0	52.0	35.3	35.9	30.1	69.5	18.3	71.6	28.4	33.8	29.6	57.1	31.1
CP50	The domestic robot plays white noise to physically prevent the microphone from recording.	53.2	29.8	47.2	32.4	51.4	31.3	49.4	26.7	67.0	21.2	66.0	20.4	48.6	29.6	51.8	28.3
CP51	The domestic robot puts a physical cover over its camera to physically prevent the camera from recording.	80.5	17.1	62.8	29.7	73.5	26.0	27.4	20.5	76.9	19.6	76.6	22.7	22.2	21.4	75.4	25.0
CP52	The domestic robot puts a physical cover over its microphone to physically prevent the microphone from recording.	50.4	26.5	42.0	26.3	52.7	25.1	44.7	23.7	67.1	15.5	65.7	19.5	35.8	29.5	56.8	26.1
CP53	The domestic robot removes its head to physically prevent the camera, microphone, and internet connection from functioning.	48.5	32.1	42.6	30.4	49.8	31.4	52.4	24.3	79.4	18.5	61.6	30.4	50.6	29.3	45.8	24.2
CP54	The domestic robot removes the LAN cable to physically prevent the internet connection.	43.0	32.5	46.5	29.5	42.0	28.1	50.8	29.7	69.8	26.9	50.9	25.4	43.6	28.3	42.3	29.4
CP55	The domestic robot removes the camera's cable to physically prevent the camera from recording.	61.8	26.9	51.4	34.4	56.8	29.1	42.6	24.9	71.0	17.8	72.8	19.5	39.6	27.2	58.8	29.1
CP56	The domestic robot removes the microphone's cable to physically prevent the microphone from recording.	76.0	16.3	69.3	28.0	69.5	22.8	41.2	27.7	76.3	21.4	81.2	17.5	23.2	20.2	73.4	22.7
CP57	The domestic robot retracts its microphone to signal that the microphone is off.	52.0	29.8	38.0	26.9	53.2	26.7	35.2	30.4	70.1	22.8	69.5	23.9	30.4	30.1	48.0	30.4
CP58	The domestic robot shows you an empty connection plug to signal that it is disconnected from the internet.	43.8	30.0	37.8	32.3	55.8	32.5	33.4	23.7	63.6	21.7	70.0	30.4	25.4	25.5	61.0	27.4
CP59	The domestic robot shows you an empty connection plug to signal that the camera is off.	46.4	21.6	37.4	28.6	49.0	30.1	46.0	27.9	67.6	20.3	60.2	25.6	34.6	26.9	44.3	26.2
CP60	The domestic robot shows you an empty connection plug to signal that the microphone is off.	53.4	19.7	48.6	29.2	55.8	21.8	32.6	20.1	59.0	17.8	64.7	26.5	24.8	24.5	60.8	19.8
CP61	The domestic robot shows you the empty connection plug to physically prevent the camera from recording.	47.4	14.2	42.4	16.6	49.8	15.7	43.2	19.5	58.8	14.7	57.2	13.0	36.3	18.4	49.0	19.5
CP62	The domestic robot shows you the empty connection plug to physically prevent the microphone from recording.	53.0	26.9	51.2	27.3	58.0	25.3	44.1	25.9	65.8	21.4	69.1	26.1	28.4	27.1	61.8	27.0
CP63	The domestic robot transforms into a different shape to signal that the camera, microphone, and internet connection are deactivated.	53.4	29.3	48.1	29.1	58.3	25.6	40.6	27.6	77.0	18.3	68.2	28.4	35.3	28.5	52.3	28.8
CP64	The domestic robot turns its camera away to physically prevent the camera from recording.	48.6	31.5	43.0	31.9	49.8	28.7	45.5	23.5	70.0	18.5	71.6	19.1	41.6	29.1	54.0	29.8
CP65	The domestic robot turns off its screen to signal that the camera, microphone, and internet connection are deactivated.	47.4	26.2	35.5	27.3	48.2	28.9	31.4	21.5	63.2	17.4	70.0	17.7	20.0	18.2	47.4	24.5
CP66	The domestic robot uses a hand gesture to signal that it is disconnected from the internet.	51.4	30.4	50.0	35.7	56.9	28.1	30.1	28.6	61.2	28.3	71.4	18.5	13.5	14.7	55.8	28.3
CP67	The domestic robot uses a hand gesture to signal that the camera is off.	40.9	30.3	43.2	31.7	51.1	27.5	27.5	24.0	55.1	27.3	56.2	25.3	18.6	19.4	47.4	30.3
CP68	The domestic robot uses a hand gesture to signal that the microphone is off.	64.4	24.8	55.5	27.1	60.6	26.0	36.8	29.9	69.2	23.1	70.8	28.3	15.5	12.4	62.9	28.2
CP69	The domestic robot uses a physical switch to physically prevent the camera from recording.	72.2	16.7	66.4	25.8	70.2	27.5	44.0	31.4	76.0	24.6	76.0	23.3	26.8	24.7	68.2	24.9
CP70	The domestic robot uses a physical switch to physically prevent the internet connection.	55.1	25.4	55.2	28.3	55.1	28.6	40.6	22.6	67.0	16.3	70.1	19.8	35.2	27.0	61.9	21.2
CP71	The domestic robot uses a physical switch to physically prevent the microphone from recording.	50.8	28.5	49.8	29.0	46.2	26.0	34.9	19.3	60.8	22.5	70.5	21.3	28.5	24.4	60.6	28.7
CP72	The domestic robot uses its voice to signal that it is disconnected from the internet.	44.8	29.0	40.4	31.6	55.2	26.1	42.2	23.5	70.7	14.3	76.2	19.2	37.8	27.0	52.3	25.0
CP73	The domestic robot uses its voice to signal that the camera is off.	37.2	27.0	33.2	27.3	38.8	31.1	44.6	25.8	67.4	19.3	73.8	20.4	33.2	27.1	44.6	30.6
CP74	The domestic robot uses its voice to signal that the microphone is off.	50.4	26.3	45.4	28.4	57.0	26.1	40.0	27.2	59.9	27.2	76.8	22.3	32.6	22.1	52.6	27.0
CP75	The domestic robot uses light feedback to signal that it is disconnected from the internet.	47.6	34.4	38.8	31.9	47.9	33.6	31.6	25.2	58.9	27.5	62.2	29.9	29.8	26.4	56.2	34.9
CP76	The domestic robot uses light feedback to signal that the camera is off.	43.6	21.4	40.0	22.9	48.8	26.1	52.4	25.9	61.3	23.8	67.2	25.1	37.0	22.8	47.1	23.2
CP77	The domestic robot uses light feedback to signal that the microphone is off.	54.4	27.3	50.4	28.2	56.6	26.5	29.2	24.2	60.7	23.5	73.4	22.8	19.5	23.7	56.4	27.8
CP78	The domestic robot uses projection to signal that it is disconnected from the internet.	48.8	25.9	49.4	28.8	59.0	28.7	38.0	28.7	63.0	28.3	67.8	24.1	25.6	23.6	60.8	20.9
CP79	The domestic robot uses projection to signal that the camera is off.	54.3	29.1	49.8	25.8	53.6	30.0	31.7	27.2	66.3	23.9	68.5	21.8	24.2	18.1	49.2	28.3
CP80	The domestic robot uses projection to signal that the microphone is off.	63.4	21.9	47.0	27.2	62.5	21.8	37.6	19.5	66.6	16.1	65.7	19.3	35.7	23.8	61.2	22.6
CP81	The domestic robot uses the smart lights in your home to signal that it is disconnected from the internet.	59.0	30.8	53.2	33.4	61.5	26.8	47.5	30.1	74.5	25.3	73.4	26.1	31.9	31.2	61.8	28.1
CP82	The domestic robot uses the smart lights in your home to signal that the camera is off.	38.8	23.9	35.2	21.9	40.2	26.3	47.6	23.6	60.0	23.6	52.0	25.7	44.6	21.4	40.0	26.0
CP83	The domestic robot uses the smart lights in your home to signal that the microphone is off.	58.2	32.4	48.2	33.9	54.6	28.7	41.0	31.8	66.6	26.9	69.9	27.5	28.8	29.0	52.2	29.1
CP84	The domestic robot waves a banner to signal that it is disconnected from the internet.	45.0	27.0	42.7	27.8	48.0	29.4	36.9	22.0	74.0	18.6	69.8	22.8	33.4	21.7	43.5	23.5
CP85	The domestic robot waves a banner to signal that the camera is off.	61.8	22.4	48.2	24.4	62.2	22.6	37.9	22.7	64.6	27.1	71.4	22.4	27.3	22.7	52.8	26.3
CP86	The domestic robot waves a banner to signal that the microphone is off.	45.8	23.6	42.9	25.3	45.9	21.5	39.2	22.0	59.6	21.3	74.0	14.1	31.0	30.0	45.6	21.1

A.3 Survey on Communication Patterns

Immerse yourself in the following situation: You have a domestic robot at home that provides entertainment and supports you with daily chores. While you appreciate the domestic robot for the convenience it provides, in some cases, you want privacy. For that, the robot uses a communication pattern to show you that your privacy is protected. Your robot does the following: *[Communication pattern.]* Please indicate to which extent you agree with the following statements:

1. This communication pattern protects my privacy very well. (Slider)

2. When the robot uses this communication pattern, I very much trust that the functionality is deactivated. (Slider)

3. This communication pattern is very effective. (Slider)

4. This communication pattern is very intrusive. (Slider)

5. This communication pattern is very noticeable. (Slider)

6. This communication pattern is very understandable. (Slider)

7. Put the slider all the way to the right side. (Attention check)

8. This communication pattern is very disturbing. (Slider)

9. I very much like my domestic robot to use this communication pattern. (Slider)

10. If you have any additional feedback, please let us know here. (Free text)
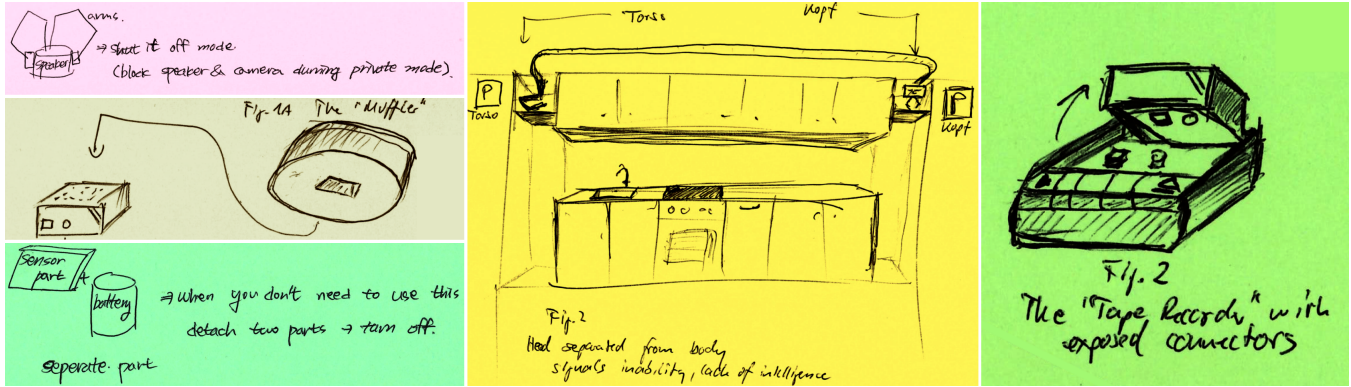
Figure 4: Examples of sketches our participants created in the focus groups.

Exploring Expandable-Grid Designs to Make iOS App Privacy Labels More Usable

Shikun Zhang
Carnegie Mellon University

Lily Klucinec
Carnegie Mellon University

Kyerra Norton
Washington University in St. Louis

Norman Sadeh
Carnegie Mellon University

Lorrie Faith Cranor
Carnegie Mellon University

Abstract

People value their privacy but often lack the time to read privacy policies. This issue is exacerbated in the context of mobile apps, given the variety of data they collect and limited screen space for disclosures. Privacy nutrition labels have been proposed to convey data practices to users succinctly, obviating the need for them to read a full privacy policy. In fall 2020, Apple introduced privacy labels for mobile apps, but research has shown that these labels are ineffective, partly due to their complexity, confusing terminology, and suboptimal information structure. We propose a new design for mobile app privacy labels that addresses information layout challenges by representing data collection and use in a color-coded, expandable grid format. We conducted a between-subjects user study with 200 Prolific participants to compare user performance when viewing our new label against the current iOS label. Our findings suggest that our design significantly improves users' ability to answer key privacy questions and reduces the time required for them to do so.

1 Introduction

Privacy policies have long been criticized for their complexity and lack of usability [32]. In response to these challenges, standardized and concise privacy nutrition labels have emerged as a potential solution to help users better understand the privacy practices of both websites and mobile apps [19–21]. Usable privacy nutrition labels can not only aid lay users' understanding of how their personal data is used, but also serve as valuable tools for privacy advocates and reg-

ulators, functioning as clear points of reference for assessing privacy practices and a foundation to enforce transparent and fair privacy regulations. Prior studies have shed light on the challenges and user frustrations associated with the existing iOS and Google privacy labels, particularly when it comes to label terminology and information layout [10, 25–27, 44, 45].

To address the information layout challenges faced by current iOS privacy labels, we built on prior research on privacy labels and access control interface design to develop and iteratively refine a prototype expandable-grid [38] privacy label that represents all iOS data categories and purposes in a color-coded compact format.

To compare our prototype labels with the existing labels, we conducted a between-subjects survey study with 200 Prolific participants. The main goals of this survey were to compare label comprehension between the existing iOS privacy labels (control condition) and our proposed label design (treatment condition), as well as explore what components contribute positively and negatively to the usability of both designs. We asked survey participants to look at the privacy labels for two existing apps with different label content and to answer comprehension questions based on the information presented in the labels. Additionally, we asked participants to provide the reasoning for their answers, which allowed us to qualitatively code their responses for sources of confusion.

Our work explores the following research questions:

RQ1: Does the proposed iOS privacy label design aid in user comprehension of iOS app data practices?

RQ2: Is the proposed iOS privacy label design effective in decreasing the time it takes for users to answer questions about mobile app data practices?

RQ3: Which elements of the existing and proposed iOS labels are most conducive or disruptive to user comprehension?

Our contributions include:

- A proposed design for an expandable-grid-based privacy label to communicate iOS app data practices.

- An empirical between-subjects study showing that the proposed design improves users' ability to answer key privacy questions and reduces the time taken to do so.

- Identification of key areas for further improvement of privacy label designs.

2   Background and Related Work

The advent of smartphones has significantly expanded the realm of mobile data processing, offering convenience and productivity to billions of users worldwide. With an increasingly diverse set of sensors and constant proximity to users, consumers are growing increasingly concerned about privacy issues associated with their mobile devices [4, 11]. The major mobile app stores have implemented and refined permission interfaces and privacy controls over the years, and have recently introduced mobile app privacy labels. In this section, we review research on mobile app privacy, privacy notices and nutrition labels, the usability of mobile app privacy labels, and tabular and grid interfaces that inspired our prototype label design.

2.1   Mobile App Privacy

Mobile devices can collect diverse and sensitive data about users, including but not limited to their location, contacts, health data, and photos. When the iPhone was first introduced in 2007, there were no permission settings until three years later [6]. Starting with the location permission, new permission settings were introduced [7]. Currently, the prevalent method of presenting privacy information and seeking consent for app permissions management systems on Android and iOS is the "ask on first use" approach, functioning as both a notice and choice mechanism. In addition, research has demonstrated the considerable influence of privacy nudges on users [1, 3, 17], and iOS added "Do you want to continue allowing this?" nudges, aimed at alerting users about background data collection.

As the number of apps grows and each app potentially requires multiple permissions, managing each and every privacy permission places an overwhelming burden on users. Recent studies highlight these usability challenges and propose the concept of "privacy assistants." These assistants can inform users about sensitive data practices and assist users in configuring privacy settings [9, 12]. Assistants can also leverage machine learning models of individual privacy preferences to further reduce user burden for privacy management [28, 30, 31, 41, 42].

2.2   Privacy Notices and Nutrition Labels

Privacy policies are the de facto standard for informing consumers about data practices, yet research has shown that these policies are prohibitively long and difficult to read [5, 32, 40]. Privacy nutrition labels were first developed by Kelley et al. as a way of addressing these issues by providing consumers with succinct descriptions of key data practices, similar to FDA food and drug labels [16]. Kelley et al. developed website privacy labels and showed they made disclosures easier to understand and reduced the amount of time people need to answer typical privacy questions [19, 20]. Later Kelley and collaborators proposed mobile app privacy labels and reported on a study suggesting that the labels would help smartphone users make better informed privacy decisions when considering apps to install on their devices [21].

In 2020, Apple introduced its own mobile app privacy labels (shown in Figure 1) and started requiring app developers to provide labels for new apps published in the iOS app store. In 2021, Google followed suit with its own variation of mobile app labels for the Google Play store.
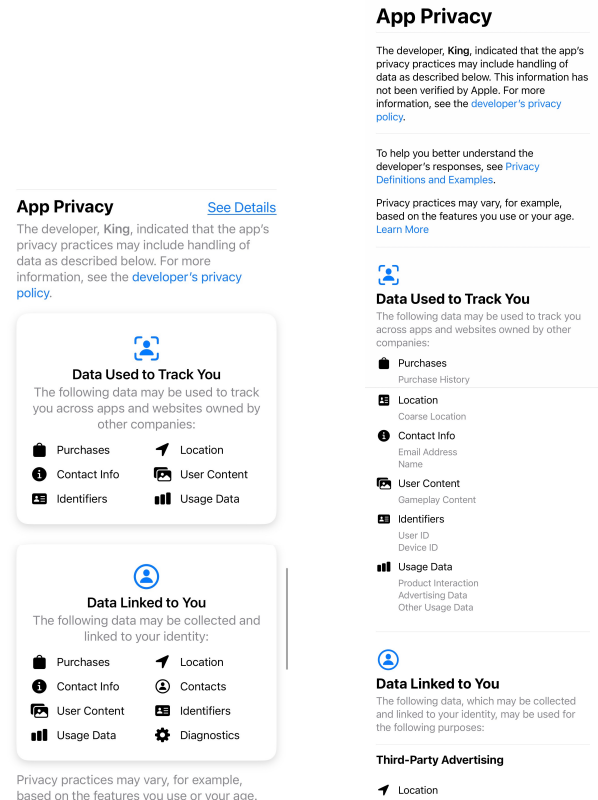
(a) Compact privacy label

(b) Detailed privacy label (partially shown)

Figure 1: Existing compact and detailed Apple Privacy Labels as found in the App Store in iOS 16.6 for the Candy Crush app. Users can click on "See Details" in the compact label to see the detailed privacy label.

2.3 Usability of Mobile App Privacy Labels

The privacy labels in the iOS App Store and Google Play Store have been widely criticized. Studies have shown that labels suffer from accuracy problems [23, 25–27], few users are aware of and use the labels, and those who try to use them find them confusing. Zhang et al. reported on a detailed analysis of iOS privacy labels, looking at the extent to which users were aware of their existence and able to use them effectively. This study revealed a number of shortcomings, including confusing label terminology (e.g., unconventional use of terms like "tracking"), confusing information organization, label complexity, and a disconnect between the labels and privacy controls made available to users [44]. This work highlighted the need to better structure label content, which is the focus of the present paper. Android data safety labels are formatted differently than iOS privacy labels and include additional information about app security. However, they suffer from similar problems as iOS labels, including confusing terminology and a complex and confusing structure [10, 29]. The diverse data practices of mobile apps pose a challenge in summarizing relevant information into an easy-to-understand format.

To make matters worse, studies have shown that despite their complexity, existing iOS and Android privacy labels may address only about half of the privacy questions typical mobile app users have [45]. However, despite recent progress towards the development of automated tools to answer users' questions by analyzing the text of privacy policies, privacy Q&A assistants are far from fully accurate [18, 35, 36]. Moreover, effective use of privacy Q&A assistants in their current state presupposes that users can both identify and articulate meaningful privacy questions. Conversely, privacy labels offer users answers to a plethora of likely questions without necessitating users to generate or articulate them independently.

2.4 Tabular and Grid Interfaces

Tables and grids, familiar to most people, present data in a concise and structured manner. Tables can typically be scanned quickly and allow for easy side-by-side comparison. They have been shown to be an effective mechanism for organizing information found in privacy policies. In particular, Kelley et al. compared tabular interfaces for website privacy policies with short- and long-text interfaces and found that people preferred the tabular format. They found their tables, which showed data types in rows and data uses in columns, were easy for study participants to use when scanning for information and comparing policies [20]. Researchers who designed and evaluated standardized financial privacy notices also found that consumers responded positively to a tabular approach [24]. In addition, tabular approaches have been used for IoT security and privacy labels [13].

Grids have been used in the design of access-control interfaces. Reeder et al. deployed a grid interface to compactly represent what access each user and group has for each file and folder in a file system. As users are often members of groups and files are often members of folders, they developed an *expandable-grid* interface that could display a grid of folders and groups, with the ability to expand any folder to show its files or expand any group to show its users. They used green and red colored cells in the grid to indicate that users were allowed or denied access to a particular file. When access permissions were the same for all files in a folder or all users in a group, the green and red colors were used on the folder or group cells. However, when permissions varied for different users in a group or different files in a folder, a yellow cell was used to indicate that expansion was needed to view detailed access permissions. Reeder et al. demonstrated that the expandable-grid approach was more effective than the traditional Windows XP access control system in making users aware of file permissions and allowing them to adjust access control settings. [37, 38]. Reeder et al. also used expandable grids in the design of a privacy label but found less success, largely due to their attempts to represent three dimensions in a two-dimensional space without using color. They offered a number of recommendations for future designers who want to use expandable grids, including representing only one dimension per axis and using short, understandable terms [39]. We leveraged expandable grids in our interface, benefiting from the lessons learned in past work.

3 Designing a New Privacy Label

Our focus in redesigning iOS app privacy labels is to improve their information layout. Both Apple and Google adopt a layered approach, offering a compact version that users can click through to get full details. But in both cases the compact version provides only minimal information, and the full version is difficult to navigate, potentially overwhelming users. Navigating the full iOS label involves extensive scrolling, and users often fail to recognize that it is a linear representation of a matrix of data types and purposes [44]. Google attempted to manage some of the complexity of the full version with an accordion interface, but users who want a full understanding of a policy must individually expand every line of the accordion, with no way to quickly scan to determine whether the app engages in a particular data practice [29].

The core principles guiding our iterative design approach were as follows: maintain a compact format suitable for mobile screens, structure the label in a more intuitive and user-friendly manner, and incorporate interactive elements to enhance user engagement and comprehension. We did not address the confusing terminology in this redesign as it requires a separate and systematic approach to identify more usable privacy terms, which is beyond the scope of this work.

3.1 Adopting an Expandable-Grid Structure

In iOS privacy labels, data practices are described along three dimensions: the data type being collected, the purpose for which that data type is collected, and whether the data being collected is linked to the user or used to track the user. In contrast to prior work by Reeder et al. [39], whose attempts to re-organize three-dimensional privacy policy data along two dimensions produced mixed results, we opted to use color to represent one dimension. We introduced a simple color scheme to represent whether data is linked to the user. We represent purpose and data type using the X and Y axis of the grid, respectively. We observed that whether data is used to track the user is actually a purpose, and therefore fold that into the purpose dimension. With 14 types of data and 33 subcategories present in Apple's privacy labels, accommodating all of them on a small mobile screen is difficult. Leveraging the inherently hierarchical relationship between data categories and subcategories (e.g., "email address" being a subcategory of "contact info"), we opted for an expandable grid format. Initially, users only see the 14 top-level data categories. Upon expanding a row associated with one of these top-level categories, the underlying subcategories of data types are revealed (see Figure 2a for an example of an expanded row). Our current label design does not include column expansion. We use color to indicate linked versus not-linked practices associated with subcategories of data, as further detailed below.

3.2 A Simple Color Scheme

In Apple's privacy labels each category of collected data may be linked ("Data Linked to You") or not linked to the user ("Data Not Linked to You"). This distinction can be captured with two colors. We use red when the collected data is linked to the user (the more privacy-invasive option), and blue, a more calming color, when the collected data is not linked to the user (the less privacy-invasive option). Entries in grey represent data types that are not collected at all.

As part of our design, we wanted to provide a summary of data practices for all the sub-categories beneath a top-level category that had not been expanded in the grid. We opted for a simple design that highlights privacy invasive practices. In this design we have five possible colors for a top-level data category with multiple underlying data types. These colors are explained in a legend accompanying our tabular format (see Figure 3). Grey indicates that no data is collected. Red indicates that a data category and any underlying sub-categories are collected and linked to the user. Dark blue indicates that a data category and any underlying sub-categories are collected but in a manner that is not linked to the user. We also introduced two additional colors to represent situations where sub-categories may be collected and used in heterogeneous ways. Salmon indicates that a subset of the underlying data types are used in a manner that is linked to the user, thereby

highlighting the existence of a privacy-invasive practice for at least one of the underlying data types (but not all). Salmon is used independently of whether some of the other underlying data types are blue or grey. The goal is simply to highlight the existence of a privacy invasive practice while also indicating that not all underlying data types are linked to the user. The light blue color is used to indicate that, while only some sub-categories of data are collected, none are linked to the user. Our salmon and light blue shades are somewhat similar in meaning to the yellow color used by Reeder et al. to represent user groups or folders in a file system with heterogeneous access permissions [38].

We considered a number of possible options including various shades of blue, red, and purple reflecting the mix of red, blue, and grey cells in underlying sub-categories. We experimented with dynamic colors based on the number of data types present and explored designs with square cells split diagonally to represent linked and unlinked data subcategories. Additionally, we considered numbers inside squares to indicate the number of underlying sub-categories. However, we opted against these options for accessibility and clarity. Our more complex designs still required expansion to understand which sub-categories were present and thus there is limited gain from such added complexity. The light colors in our design serve as a cue for row expansion.

3.3 Adding Interactive Elements

We designed the grid to be expandable so that users could tap on a row label or chevron to expand a row or collapse a row already expanded. In addition, users can tap on individual cells in our table to access more detailed information about the meaning of each cell, the data it corresponds to, and the practices it describes, including how many subcategories of data it represents and the purpose of data collection associated with this particular entry (see Figure 2c for an example).

When Apple first introduced its privacy labels, they were static notices that lacked any interactive features. There was only a "See Details" link at the top right corner of the compact label (Figure 1a), linking to the detailed label (Figure 1b). In prior studies of iOS privacy labels [44], users expressed a desire for more interactive labels. Later, Apple changed its labels so that users who tap on each section within the compact labels are brought to the corresponding section of the detailed privacy label. However, the iOS labels still do not offer a direct link to definitions of terms used (a list of definitions is available only in the detailed view after users tap on "See Details"). To make definitions of terms more accessible, we placed information icons next to relevant terms; tapping one of these icons triggers a pop-up with a definition of the term, as shown in Figure 2b. To make it easier to expand the grid and access the popovers on a small screen, we designed the interface so that a tap anywhere near a row label expands the row and a tap anywhere in a cell triggers the

popover. Tapping outside the popover or on another element closes it. This seemed to work well for our pilot participants.

As our legend does not always fit on the same screen as the label, we incorporated a hyperlink within the table. This hyperlink ("What do the colors and symbols mean?") enables users to readily jump to the legend for details. Figure 4 shows our label on two different screen sizes.

(a) Clicking on a row

(b) Clicking on the "i" icon

(c) Clicking on a cell
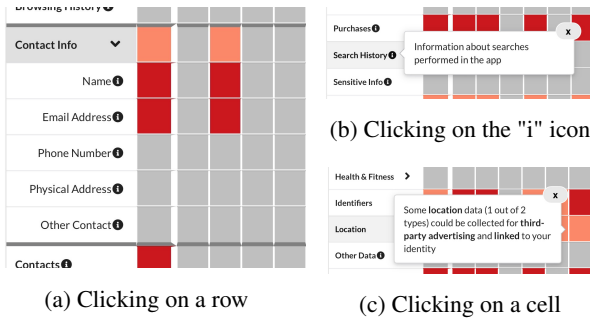
Figure 2: Interactive elements in treatment labels
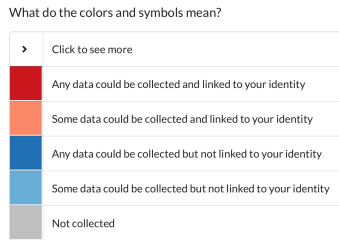
Figure 3: Legend used in treatment labels

3.4   Interview Pilot

We conducted two rounds of small-scale semi-structured interviews to help gain rich insights into the strengths and weaknesses of our prototypes. All pilot participants were assigned to view either the iOS privacy labels or our prototype labels in a round-robin fashion. These interviews were conducted over Zoom on their iPhone and participants shared their screens while interacting with the privacy labels on a mobile website that we created. This enabled us to record what actions they took with the label while answering our questions.

We asked participants about their prior experience with privacy labels and whether privacy ever influenced their decision to stop using an app. Then, we sent participants links in Zoom to open the label on their iPhone. The second section of the interview assessed participants' ability to accurately answer questions based on label information. Afterwards, we asked participants about the definitions of terms used on the labels. Finally, we asked participants to identify helpful or unhelpful aspects of the labels and provide additional feedback.
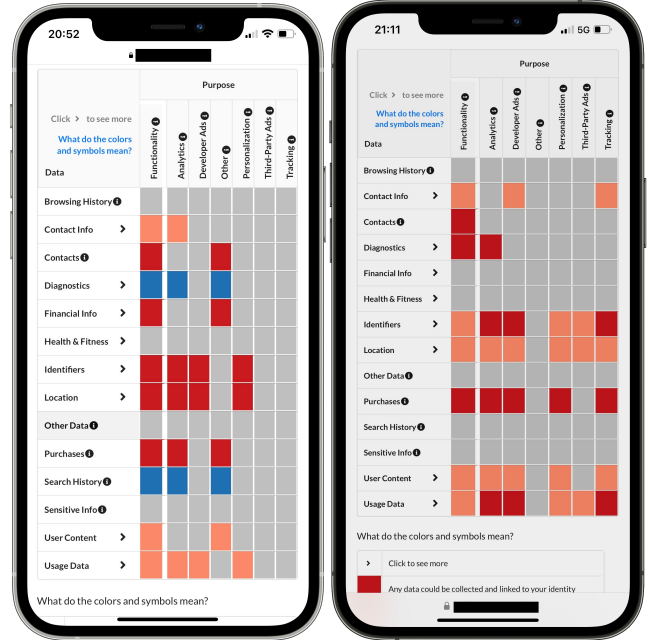
Figure 4: Treatment labels on two different screen sizes: Venmo label on an iPhone 12 Pro (left) and Candy Crush label on an iPhone 12 Pro Max (right)

Insights from the interview phase informed some modifications to the label and the development of our survey protocol. For example, our early label design hid some of the less common data categories under a "see more" row, but we found this confused pilot participants so we showed rows for all categories. In addition, the early version ordered data types and purposes by frequency in the App Store. However, pilot participants did not understand this so we switched to alphabetical order.

4   Methods

In this section, we describe our study design. We describe our recruitment process, survey procedure, survey pilots, thematic analysis, and limitations of this study.

Ethical considerations.   Our interview pilot, survey pilots, and main survey were reviewed and approved by the Carnegie Mellon University Institutional Review Board. All study participants completed online consent forms.

4.1   Recruitment

We recruited participants on the Prolific[1] research participant recruitment platform who were iPhone users running iOS 14

---

[1] https://www.prolific.com/

or a newer version of the operating system, and thus had iOS labels available on their phones. The number of participants was determined after performing a power analysis as detailed in Section 4.3. We recruited participants who were over 18, fluent in English, and residing in the United States. We required participants to take the survey on their computers while viewing the privacy labels on their iPhones. All of these criteria were checked using Prolific's built-in screening capabilities to prevent ineligible participants from accessing the survey. Additionally, we set parameters on Prolific to create a balanced sample in terms of gender. We did not set any screening criteria for other demographic factors and we did not collect any demographic data in our survey beyond what was collected automatically by Prolific. Participants were paid $5 for successfully completing the survey.

4.2 Survey Design

We used a between-subjects survey design where participants were randomly assigned to either view Apple's privacy labels or our prototype labels. The survey consisted of four main parts: general questions about privacy and privacy labels, questions about the information found in the privacy labels for two different apps, questions about terms used in privacy labels, and feedback about the labels they were shown.

4.2.1 Study Apps

We selected two popular apps that represent significantly different types of privacy labels: Candy Crush Saga and Venmo. Candy Crush Saga has a "Data Used to Track You" section, whereas Venmo does not and instead has a "Data Not Linked to You" section. Both apps have "Data Linked to You" sections. Neither app has all three sections since iOS privacy labels with three sections are less common than those with two [2]. See Figure 1a for an example of Candy Crush's compact iOS privacy label. In our label design (shown in Figure 4), the Venmo label has blue squares to represent data that is collected but not linked to identity and red and salmon squares to represent data linked to identity, while the Candy Crush label only has red and salmon squares that represent data collected and linked to identity.

4.2.2 Survey Procedure

The survey began with general questions about privacy and privacy labels. Next, we prompted participants to use their iPhone's camera to scan a dynamically generated QR code, which encoded their Prolific ID and sent them to a specific privacy label based on their condition. We opted to show the labels on participants' phones instead of on computers to ensure that participants interact with the privacy labels in a more ecologically valid setting. Upon scanning the QR code, participants were directed to a webpage simulating the Apple

App Store environment for either the Candy Crush or Venmo app. Full webpage representations shown to participants can be found in Appendix C. Participants then responded to six comprehension questions related to the app privacy label they were viewing. They were encouraged to interact with the label on their iPhone while answering the questions. Then, they scanned another QR code to view the second app and answer the same set of six questions for the second app label. The study website used Javascript to record participants' actions, including scrolls, taps, and associated timestamps. In addition, the study website also checked the browser user agent string, confirming that participants were indeed viewing the labels on an iPhone running iOS 14 or above.

4.2.3 Conditions

Half of the participants were randomly assigned to the control condition ($N = 100$) (viewing the current iOS privacy labels) and the other half to the treatment condition (viewing our prototype labels). In both conditions, participants saw the corresponding labels for two apps. Within each condition, we also randomized the order in which participants encountered each app (Venmo first or Candy Crush first). As a result, participants were randomly assigned to one of four possible groups in Qualtrics: `Control Candy`, `Control Venmo`, `Treatment Candy`, and `Treatment Venmo`, each *group* comprising approximately 50 participants. For instance, a participant in the `Control Venmo` group was in the control condition and saw Apple's Venmo label first and then Apple's Candy Crush label. This allowed us to both compare the treatment and control labels as well as see whether the order in which the apps were viewed affected participant performance.

4.2.4 Comprehension Questions

We asked six multiple-choice comprehension questions about each app's privacy label, totaling twelve questions. The questions were designed to elicit all potential types of interactions users could have with the labels in both conditions. These questions represent typical user privacy questions that can be answered using the labels, i.e., questions about types of data collected and purpose. Prior research [45] found that about 30% of contextualized user questions about mobile apps are related to data types and purpose of data collection. They include questions such as whether an app might collect photos and videos, or whether diagnostic data might be linked to a user's identity. Each question had different answers for each of the two apps of interest, preventing participants from using the same answers for both questions. Table 1 shows the questions that were asked, their respective question category, the correct answers for each label, and the actions participants would need to take to find the correct answer for each condition. We also asked participants to provide open text explanations for each multiple-choice question. In Section 5, we denote the

Question			Answer		Action	
#	Text	Category	Venmo	Candy	Treatment	Control
Q1	Does this app collect data for Analytics purposes and, if so, what data? (Select all that apply)	Any data type for one purpose	Contact info, Diagnostics, Identifiers, Location, Purchases, Usage data	Diagnostics, Identifiers, Location, Purchases, User content, Usage data	Look down a column	See details and find all data types under a purpose
Q2	Does this app collect location data for Third-Party Ads purposes?	One data subtype for one purpose	No	Coarse location	Expand row and look at a cell	See details and find a data type under a purpose
Q3	Does this app collect Photo and Video data and, if so, for what purpose(s)? (Select all that apply)	One data subtype for any purpose	(App) Functionality, Other	(App) Functionality	Expand row and look at a row	See details and find a data subtype under all purposes
Q4	Does this app collect Purchases data and, if so, for what purpose(s)? (Select all that apply)	One data type for any purpose	(App) Functionality, Analytics, Other	(App) Functionality, Analytics, Developer Ads/Advertising, Other, Tracking	Look at a row	Find a data type in compact view, see details, find a data type under all purposes
Q5	Does this app link Diagnostics data to your identity?	One data type is linked or not	No	Yes	Look at a row's color + legend	Find a data type in compact view
Q6	Does this app collect data for Tracking purposes, and if so, what data? (Select all that apply)	All data types for one purpose	No	Contact Info, Identifiers, Location, Purchases, User content, Usage data	Look down a column	Find all data types in compact view

Table 1: Questions used in the survey, corresponding correct answers for each app, and the action needed for participants in the treatment and control conditions to answer each question correctly

questions for Candy Crush and Venmo as "CQ1–CQ6" and "VQ1–VQ6" respectively.

4.3 Survey Pilots

We conducted two rounds of survey pilots to make sure the survey protocol (including the server hosting and recording participant actions and the Qualtrics survey flow) worked as intended and to collect data for use in our power analysis.

We conducted the first pilot survey with 40 participants on Prolific under the same recruitment criteria as our main survey. We conducted an a priori power analysis using G*Power. T-test was chosen as the test family, and the Wilcoxon-Mann-Whitney test of the mean accuracies between two conditions was selected. We chose the usual alpha level of 0.05 and the most common beta value of 0.2 (indicating a power of 0.8) [33] to calculate the minimum sample size necessary for detecting the expected effect as estimated by the pilot sample. The detailed results of the power analysis for each of the 12 questions (Table 4) can be found in the Appendix. This analysis ensured that our study (with 100 participants per condition) was adequately powered to detect significant differences in the accuracy between the control and treatment conditions for Questions Q1, Q2, Q4, and Q6. As the effect sizes were small for Q3 and Q5, the power analysis suggested we would need a much larger sample size to detect significant differences between the control and treatment. We selected our sample size based on the other questions, but we still included Q3 and Q5 to observe the effectiveness of design mechanisms (e.g., color, row expansion) noted in Table 1.

4.4 Thematic Analysis

In our thematic analysis of the open text explanations provided by participants for the comprehension questions, we employed both inductive and deductive coding methods [8, 15].

To ensure internal reliability, three of the authors participated as coders and inductively coded the responses. Each response was coded by two authors. Our coding process included the following steps: three authors read through the responses to develop a set of codes. The first author reviewed all responses. The second author focused on the treatment responses, while the third author focused on the control responses. After developing these initial codes, the authors discussed the definitions and adjusted the codes based on their discussion. As our interest was primarily in the reason for incorrect answers, the three authors independently coded the explanations for a set of 812 incorrect answers and compared their coding. Any disagreements were resolved, which resulted in adjusted definitions. Finally, the authors proceeded to code participant responses using the revised codebook and resolved all conflicts.

4.5 Limitations

We enrolled participants whose iPhones were running iOS 14 or above because the iOS privacy labels were only available for those users. Additionally, our participant pool was limited to Prolific users in the United States who were proficient in English. We focused only on participants from one region, the United States, because App Store interfaces and available apps vary by region. This allows us to ensure the consistency of our simulated presentation of the privacy labels in the App Store, aligning with participants' prior experiences and mitigating the introduction of unaccounted variables. However, our results may not generalize to users in other regions of the world. Subsequent investigations could delve into the potential influence of using various languages in the labels or broaden the scope to encompass additional cultural variables. Moreover, our study focused on just two apps, Candy Crush and Venmo, and their corresponding iOS privacy labels. Users might have different experiences using other app labels or after becoming more acquainted with the labels over time.

Finally, our study focused on use of the labels by participants assigned to use them and may not fully reflect the experience of users who are motivated to review labels of apps they actually use or are considering using.

5 Results

We first present information about our participants, followed by results on accuracy, errors, time answering questions, perceived confidence, learning effect, interaction with treatment labels, and understanding of iOS label section headers.

5.1 Participants

We manually removed 15 participants due to low-quality free-text responses, lacking necessary interactions (e.g., scrolls, visiting both app labels), or answering too quickly. We analyzed the demographic information provided by Prolific. Our sample is balanced with 100 male and 100 female participants, all of whom met our specified criteria: fluent in English, iPhone users, and residing in the United States. Further details regarding the distribution of participant ages and ethnicities can be found in the Appendix. Our participants are experienced Prolific users with an average total approval count [2] of 1262 ± 1168 tasks. Our minimum approval count is 16 with 3 participants having less than 50 approvals. The median completion time was 21.2 minutes.

5.2 Accuracy Analysis

Our survey included six comprehension questions (Table 1) for each of two apps. We assessed the performance of participants in both the control and treatment conditions based on the accuracy of answers they provided.

5.2.1 Significant Differences in Half of the Questions

Figure 5a shows the accuracy percentages (the proportions of correct responses) for each of the 12 questions in each condition. We observed that the treatment group outperformed the control group in 9 out of the 12 questions. In one question (CQ6), both conditions had the same accuracy percentage. However, in two instances (CQ2 and VQ5), the control group outperformed the treatment group. To assess the statistical significance of these differences, we conducted pairwise Fisher's exact tests between the control and treatment groups for all 12 questions; the Holm-Bonferroni corrected p-values for these tests are also marked in Figure 5a. We obtained statistically significant results for half of the questions: for five of these questions the treatment group outperformed the control group and for one question the control group outperformed the treatment group.

---

[2]The total approval count represents an individual participant's number of approved submissions for tasks on Prolific.

5.2.2 Treatment Outperforms Control When Data Collection Is Absent

In the case of questions VQ2 and VQ6, which require participants to determine that Venmo does not engage in the data practices discussed in these questions, the treatment condition performed significantly better than the control group. For VQ2, this improvement arises because the treatment condition clearly indicates the absence of data collection with a gray-colored square, whereas participants in the control condition need to inspect the relevant sections to discern this absence i.e., participants need to search for "coarse location" within the detailed label under the "third-party advertising" purpose category and recognize that it is not there. This distinction becomes particularly evident when comparing the same question between Candy Crush (CQ2) and Venmo (VQ2) apps in Figure 5a for the control condition. In CQ2, around 90% of the participants were able to answer correctly when the data practice is there; while in VQ2, less than 40% of the control participants answer it correctly.

5.3 Error Analysis

We examined the number and type of errors made by participants in each condition, identifying common error themes across both conditions as well as errors that frequently occurred in just one of the two conditions.

5.3.1 Treatment Significantly Reduced Errors

We calculated the mean number of incorrect answers for each condition. In the treatment group, the mean number of incorrect answers was 2.68 with a standard deviation of 2.19, while the control group had a mean of 5.08 incorrect answers with a standard deviation of 2.38. We employed the Mann-Whitney U test to compare two independent groups (control and treatment) as the data is not normally distributed. The test confirmed the significant difference between the two groups (U = 2230.0, p = 4.53e-12) with a large effect of size 0.55. The treatment significantly improved on the control, reducing errors by approximately half.

5.3.2 Common Error Themes Across Conditions

We analyzed 812 explanations for incorrect answers and identified a number of common error themes for both conditions during our qualitative analysis. First, many participants were confused by the terminology used in the labels, such as conflating "identifiers" with "linked to your identity" in Q5, mixing up "contacts" and "contact info" in Q1 and Q6, and struggling to differentiate between "developer advertising" and "third-party advertising" in Q2. This confusion often led them to search in the wrong part of the label for answers. Second, some participants misunderstood the questions or provided responses based on their personal beliefs or prior knowledge

rather than the information provided in the labels. For instance, in the case of asking whether Venmo collects purchase data, one participant answered, "They do keep record of your bank account login information, routing numbers and credit cards linked to your account, but they do not disclose and[sic] information to third party social networking services." Third, some participants made accidental errors or mistakes when answering the multiple-choice questions but quickly realized and explained them in their free-text justifications. Fourth, some participants provided vague or brief justifications, making it difficult for us to pinpoint the reasons behind their error.

5.3.3 Challenges with Color Coding for Treatment

In two of the questions (CQ2 and VQ5), where the treatment condition showed worse performance compared to the control condition, errors were related to the use of color coding within the treatment labels. In the case of CQ2, the correct answer for Candy Crush is indicated by a salmon-colored square in the treatment label. Participants first need to understand that the color signifies certain sub-categories, but not all "Location" sub-categories are collected. Then participants must expand the row to know whether the salmon-colored square signifies "coarse location" or "precise location" being collected. The qualitative analysis revealed that 19 participants (60% of the incorrect participants) could not find the info or provided answers that suggested they did not expand the row. This is also consistent with the recorded taps where 18 (58% of incorrect participants) did not expand the row. The 30% error rate for this question also aligns with participants' comprehension rate of color cues, as evaluated in the treatment condition later in the survey, where 31% of participants did not seem to understand that a salmon-colored square indicates less data is collected than a red square.

VQ5 pertains to whether diagnostics data is linked or not to user identity. In the treatment condition, participants need to recognize that a blue square signifies that the data is not linked to user identity. From our qualitative analysis of participants' justification, we found that 14 participants (50% of the incorrect answers) misinterpreted the colors, and another 5 participants (18% of the incorrect answers) accidentally selected the wrong answer or immediately realized that they had selected the wrong answer as explained in their free-text justifications. We also assessed participants' ability to correctly interpret the blue color in a later question, with 80% of participants correctly interpreting the meaning of the color.

5.3.4 Incomplete Answers for the Control

We observed that a major reason for incorrect answers in the control was incomplete answers. This pattern is very evident in the case of Q1 (finding all data types for analytics purposes). For VQ1, participants need to find all data types used for analytics across sections "Data Linked to you" and "Data

Not Linked to You," which required a lot of scrolling in the control condition. No control participants answered correctly. In CQ1, where all correct answers fell under the "Data Linked to You" section, participants were more accurate, with a 45% error rate. Our qualitative analysis showed that 79% of the errors were due to participants not scrolling enough to find all the information. Another common error was to select all data types as answer choices (20% of the errors).

Furthermore, a significant drop in performance was observed in the control group when comparing CQ4 and VQ4. The sole difference between the two was that participants had to identify 3 purposes for CQ4 and 5 for VQ4. Control participants were more likely to provide incomplete answers when faced with a higher number of purposes. Conversely, treatment participants responded with high accuracy regardless of the number of purposes they had to identify.

Each accuracy question also asked participants to explain how they arrived at their answers through a free-text response. As described in Section 4.4, we thematically coded these responses. Below, we present the primary themes that emerged during this analysis, along with their respective frequencies.

5.4 Time to Answer Comprehension Questions

We computed the time spent on comprehension questions, excluding the time for free-text responses. For the control condition, the mean time was 10m59s, while for the treatment condition, it was 8m28s. Since the time spent does not follow a normal distribution, we conducted the Mann-Whitney U test, which revealed a statistically significant difference between the control and treatment conditions (U = 2202.0, p = 4.78e-07) with a large effect size of 0.56.

These findings indicate that participants in the treatment condition spent significantly less time compared to those in the control condition. Figure 5b provides a detailed breakdown of the time spent answering each of the 12 questions. Our timing data includes time for both correct and incorrect answers. When we specifically examined the time for correct answers, we found the same trends. We further conducted pairwise Mann-Whitney U tests between the control and treatment conditions for each of the 12 questions and applied Holm-Bonferroni correction to the p-values. As shown in Figure 5b, 8 out of the 12 questions produced significant results.

In all questions except one (CQ2), the control group took more time than the treatment. In VQ2 and VQ6, where the answer is "no" and thus there is no mention of that type of data collection in the control, the control took significantly more time than the treatment with a large difference.

5.5 Perceived Confidence Analysis

For each question, we also asked participants to rate their confidence in their answers on a Likert scale ranging from 1 (not at all confident) to 5 (extremely confident). The distribution

(a) Accuracy percentages by condition

(b) Time spent to answer questions
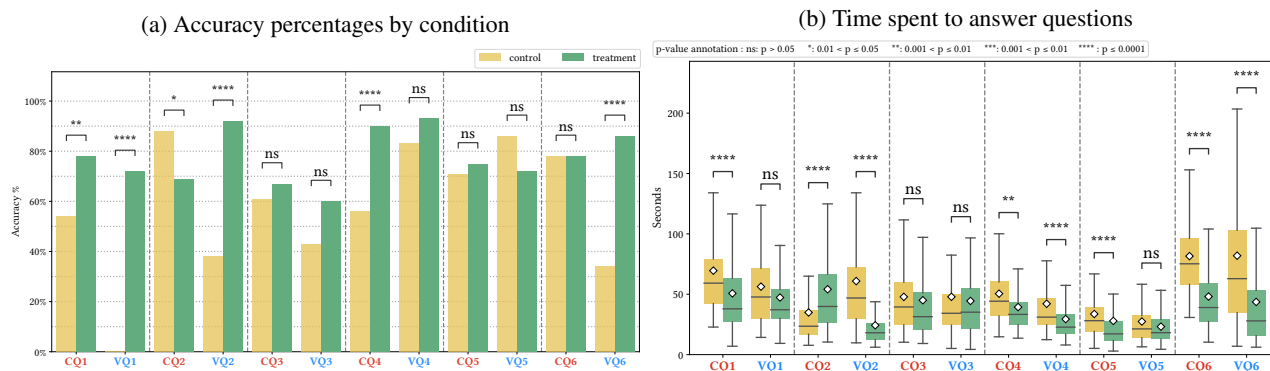
Figure 5: CQ1 denotes Question 1 for the app Candy Crush, and VQ1 denotes Question 1 for the app Venmo. All p-values adjusted by the Holm-Bonferroni method.

	Incorrect Count						Time Spent in Minutes					
	Candy 1st	Candy 2nd	p	Venmo 1st	Venmo 2nd	p	Candy 1st	Candy 2nd	p	Venmo 1st	Venmo 2nd	p
Control	2.39 ± 1.40	1.43 ± 1.38	**	3.39 ± 1.30	2.94 ± 1.41	ns	5.26 ± 1.48	5.87 ± 3.60	ns	6.89 ± 3.70	4.01 ± 1.87	****
Treatment	1.73 ± 1.30	1.12 ± 1.05	*	1.41 ± 1.44	1.10 ± 1.17	ns	5.70 ± 3.64	3.70 ± 3.73	****	4.49 ± 2.75	3.03 ± 1.60	**

Table 2: The difference in the number of incorrect answers (left) and time spent in minutes (right) among participants viewing Candy or Venmo as either their first or second app across both the control and treatment conditions. Eight one-sided Mann-Whitney U tests were conducted in total, and p-values were adjusted using the Holm-Bonferroni method.

of responses is graphed in Appendix A, Figure 6. The control condition exhibited a significant level of uncertainty regarding VQ2 and VQ6, both of which pertain to situations where the data collection mentioned in the question is absent from the label. In 9 out of 12 questions, more participants in the treatment group felt "extremely confident" in their answers compared to the control group. The mean confidence score for the control condition is 4.29 ± 0.49, while for the treatment condition, it is 4.48 ± 0.63. We also calculated the Kendall's Tau-b rank correlation between the average confidence level and the number of incorrectly answered questions across both conditions. The Kendall's Tau-b is a non-parametric measure of association that exists between two ordinal variables [22]. In the control group, Kendall's tau correlation revealed a statistically significant negative and *weak*[3] relationship between the two variables ($\tau = -0.17, p = 0.02$). For the treatment, a statistically significant negative *strong* correlation was observed between the two variables ($\tau = -0.35, p < 0.0001$). This stronger correlation in the treatment group suggests that participants in this condition were more likely to feel confident about their answers when they were indeed correct compared to the control group.

5.6 Learning Effect

To assess the presence of a potential learning effect between the first and second exposure to the labels, we further divided

each of the control and treatment conditions into two distinct groups: the "first" group consists of participants encountering the label as the first label they viewed in the study, while the "second" group comprises participants encountering the label as the second label they viewed in the study. We consider the learning effect in two dimensions: accuracy and time.

Even though we observed a decrease in errors from first to second for both apps and both conditions (Table 2), after correcting p-values for multiple tests, only the decreases in errors for Candy Crush are significant across both conditions. Table 2 shows the average time for participants in each condition for both groups. There is a significant decrease in the time needed to answer questions for the treatment group no matter which app they see first. However, the decrease in time only appears for participants answering questions for the Venmo app in the control condition.

5.7 Interaction with Treatment Labels

As noted in Section 4.2.2, we captured participants' interactions with treatment labels, including taps and scrolls.

Our analysis revealed that among 100 participants in the treatment condition, 80% of participants expanded one or more rows during the study, 68% of participants tapped at least one of the information icons to access definitions, 61% of participants tapped at least one cell, and 32% of participants tapped the hyperlink (i.e., "What do the colors and symbols mean?") located inside the table that brings them to the legend. These findings indicate that participants actively engaged with

---

[3] https://www.spss-tutorials.com/kendalls-tau/#kendalls-tau-formulas

the interactive elements incorporated into the treatment labels.

Analyzing participant interaction with information icons revealed that the purpose category "Tracking" garnered the highest number of taps at 76, followed by "Other" with 36 taps. The data type "Other" received 59 taps, while "Purchases" received 41 taps. Other information icon interactions with more than 10 taps include: "Sensitive info" (23 taps), "Other diagnostic data" (19 taps), "Third-party advertising" (14 taps), "Analytics" (13 taps), "App functionality" (13 taps), "Diagnostics–crash data" (12 taps), and "Diagnostics–performance data" (11 taps).

5.8 Participants' Understanding of iOS Privacy Label Section Headers

To briefly explore participants' understanding of the terminology used in the privacy labels, we asked participants multiple-choice questions regarding the definitions of "Data Linked to You," "Data Not Linked to You," and "Data Used to Track You." 74% of the participants correctly identified the definition of data linked to you and 49% correctly identified the definition of data not linked to you, with no significant differences between the control and the treatment conditions.

For "Data Used to Track You," only 53% of the control and 33% of the treatment were correct, showing a significant difference with a p-value of 0.016 after Bonferroni correction. One potential explanation for the treatment label's poorer performance is treatment participants were exposed to the term "Tracking" rather than "Data used to track you" in the interface, but still asked about "Data used to track you" in the survey. Another potential explanation is that the treatment label displayed "Tracking" as a purpose, alongside other purposes such as "Personalization" and "Third-Party Ads." The correct definition of tracking—"Identifiable data that is shared with third parties to personalize ads" contains words similar to these other purposes. This might have led participants to believe that tracking should be distinct from these listed purposes. To delve deeper, we observed that out of 28 treatment participants who clicked on the information icon for tracking at some point during the study, 18 answered this question correctly. In contrast only 15 out of 72 treatment group participants who did not click on the information icon for tracking were correct. This indicates that the information icon likely contributed to participants selecting the correct definition of tracking.

We also described two data collection scenarios and asked participants whether they consider each to be tracking or not: 1) app sharing your location/email address with third party advertisers, 2) app using location to show you nearby stores. For the first scenario, 81.5% correctly consider that to be tracking, but for the second scenario only 6% correctly identified that it is not tracking under Apple's definition. The responses to these scenario questions were not significantly different between control and treatment conditions.

6 Discussion

Below, we summarize the main findings of our research and discuss future possible avenues for extending this work, including addressing privacy label limitations not addressed by the proposed grid layout (e.g., confusing terminology) and opportunities to offer personalized label presentations.

6.1 What Made the Treatment Effective

Expandable grids have been evaluated in various contexts with mixed results: they were shown to be well suited for windows file permission control but less effective for P3P policies [37–39]. Our study reaffirmed the advantage of displaying the complete policy [38]. In one early design variation, we opted to display only selected rows, requiring users to click "see more" for additional content. However, many pilot interview participants missed accessing the complete content. The expandable grid format enabled us to accommodate the limited screen real estate available on mobile devices and present the entire label in a compact, organized format. In contrast, the lengthy format of the control label resulted in incomplete answers due to the need for extensive scrolling and compiling answers from multiple sections. This improvement was instrumental in helping participants answer questions such as Q4 correctly, which requires them to consider all purposes associated with the collection of a specific data type. Treatment participants could readily answer the question by inspecting a single row in the table, whereas control participants had to scroll through a number of purpose sections spanning multiple screens within the "See Details" view.

Prior research suggests that an effective approach involves developing an expandable grid representing one dimension per axis and incorporating color to represent a third dimension [38, 39]. Reeder et al. also found that juxtaposing two dimensions on a single axis was confusing to users [39]. The current full iOS privacy labels represent the two dimensions of data using a list, which did not work well with users [44]. In our design, we arranged data type and purpose along the X and Y axes, while employing color as the third dimension.

Reeder et al. also noted that despite multiple cues in the P3P Expandable Grid, 14.5% of participants did not seem to notice that they could expand the grid [39], a problem we also encountered with 20% of our participants failing to expand rows. On the other hand, Zhang et al. observed that iOS label users expected interactive privacy labels on mobile screens, and were disappointed when they could not tap on the label to access privacy choices or additional information [44]. 91% of users in our study did engage with the interactive components of our labels. This interaction seems to facilitate user comprehension and enhances usability of the labels.

Treatment participants performed significantly better and more quickly than control participants when the particular data collection practices they were looking for were not

present within the labels. In such scenarios, they could spot the gray-colored squares that effectively signaled the absence of certain data practices. This aids users in swiftly identifying apps not collecting certain types of data at a glance. Additionally, our correlation analysis indicates a notably stronger negative correlation between participants' confidence levels and errors in the treatment condition (namely, treatment participants answered more correctly and more confidently)

6.2 How To Improve the Treatment

Introducing Users to Row Expansion and Legend.   Many of the treatment errors were attributed to users not expanding rows, as noted above. We did not provide any training to help users become familiar with the labels in either condition. It would be beneficial for the interface to include a quick integrated tutorial or animated cues to help users understand the legend and the row expansion, which could improve accuracy.

Addressing Accessibility Concerns.   Considering that our treatment prototype relies on color coding, there is an accessibility issue for individuals who are color blind. To mitigate this concern, we carefully selected colors that are accessible for people with various color vision conditions except monochromacy. We recognize the limitations of relying solely on color and future research could explore the integration of dot or stripe patterns and other features to further enhance clarity and accessibility. Additionally, the use of a grid may also raise further accessibility concerns for individuals with visual impairments, including those who have low vision. These elements may be difficult to handle for screen readers, which are tools commonly used by visually impaired users. We note that the current version of the label deployed in the app store is also tedious to navigate with a screen reader. While our results suggest that our proposed design could help many users, addressing the needs of the visually impaired community when it comes to benefiting from privacy labels will require more work.

Improving Terminology.   It is also worth noting that our treatment labels did not address the issue of confusing terminology, a pain point identified by participants in prior studies [27, 44]. This decision was deliberate, because we believe that rectifying this problem necessitates a systematic and comprehensive approach to identifying more intuitive terminology. Our findings also provide further evidence of the confusion created by some of the terms used in existing privacy labels, especially when it comes to Apple's definition of tracking. Not only did participants fail to answer the questions regarding tracking correctly, their interaction with the information icons echoed the same trend. The interactive information icon for "tracking" received the highest number of taps at 76. In contrast, other purpose terms such as "analytics"

and "third-party advertising," which were also included in the comprehension questions, each received under 20 taps.

In addition, our results suggest that the terminology used to refer to some top level categories of data types is also unintuitive, with users struggling to identify the top-level data category for some data types (e.g., "Photos and Videos" falling under "User Content"). The information icons for the two "Other" terms (one for purpose and one for data type), also attracted a great number of clicks from our participants, indicating participants' need for additional information. This aligns with previous research findings [44], indicating that participants expressed confusion when encountering terms in the label associated with other data types or other purpose. Further research will be needed to address these issues.

6.3 Future Directions

Comparing App Labels.   Ultimately, we believe that privacy interfaces should be designed to empower users to readily compare the data practices associated with similar apps such as two apps in the same category. We believe that the tabular format presented in this paper will naturally lend itself to a comparison interface that can highlight cells where two apps have diverging data practices, allowing the user to quickly zoom in on key differences. Future work could also explore ways to use the proposed grid layout to highlight practices that are atypical of similar apps in the app store.

Personalized Label Presentations.   While the grid format in this study is clearly improving usability, privacy labels remain complex. A further opportunity to enhance usability might involve exploring personalized presentations of privacy labels, as has been prototyped for IoT labels [14], letting users choose which practices interest them and which they don't care about. Such an approach could also benefit from the use of machine learning to assist users in making these selections (e.g., [30, 31, 41, 43]). A tabular format similar to the one evaluated in this study could be adapted to highlight data practices of interest or highlight practices that are likely to deviate from the user's expectations (e.g., [34]).

7 Conclusion

We propose an expandable-grid-based privacy label designed to improve the usability and mobile app privacy communication over current iOS labels. Our between-subjects study with 200 Prolific participants shows significant user improvement in answering privacy questions more accurately and faster. We believe that our redesign contributes to better informing consumers about the privacy implications of their future app downloads. We hope that this research will inform the design of more effective mobile app privacy labels and the development of effective privacy labels in other domains such as websites and IoT devices.

References

[1] Hazim Almuhimedi, Florian Schaub, Norman Sadeh, Idris Adjerid, Alessandro Acquisti, Joshua Gluck, Lorrie Faith Cranor, and Yuvraj Agarwal. Your location has been shared 5,398 times! A field study on mobile app privacy nudging. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*, pages 787–796, 2015.

[2] David G. Balash, Mir Masood Ali, Xiaoyuan Wu, Chris Kanich, and Adam J. Aviv. Longitudinal analysis of privacy labels in the apple app store, 2023.

[3] Rebecca Balebako, Jaeyeon Jung, Wei Lu, Lorrie Faith Cranor, and Carolyn Nguyen. "Little brothers watching you": Raising awareness of data leaks on smartphones. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, SOUPS '13, New York, NY, USA, 2013. Association for Computing Machinery.

[4] Jan Lauren Boyles, Aaron Smith, and Mary Madden. Privacy and data management on mobile devices. *Pew Internet & American Life Project*, 4:1–19, 2012.

[5] Rex Chen, Fei Fang, Thomas Norton, Aleecia M McDonald, and Norman Sadeh. Fighting the fog: Evaluating the clarity of privacy disclosures in the age of CCPA. In *Proceedings of the 20th Workshop on Workshop on Privacy in the Electronic Society*, pages 73–102, 2021.

[6] Jacqui Cheng. Ars reviews iOS 4: What's new, notable, and what needs work. https://arstechnica.com/gadgets/2010/06/ars-reviews-ios-4-whats-new-and-notable/7/, Jun 2010.

[7] Jacqui Cheng. Review: iOS 6 gets the spit and polish treatment. https://arstechnica.com/gadgets/2012/09/review-ios-6-gets-the-spit-and-polish-treatment/, Sep 2012.

[8] Victoria Clarke, Virginia Braun, and Nikki Hayfield. Thematic analysis. In *Qualitative psychology: A practical guide to research methods*, page 248. SAGE, 3rd edition, 2015.

[9] Jessica Colnago, Yuanyuan Feng, Tharangini Palanivel, Sarah Pearman, Megan Ung, Alessandro Acquisti, Lorrie Faith Cranor, and Norman Sadeh. Informing the design of a personalized privacy assistant for the internet of things. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[10] Lorrie Faith Cranor. Mobile-app privacy nutrition labels missing key ingredients for success. *Commun. ACM*, 65(11):26–28, oct 2022.

[11] Cybersecurity and Infrastructure Security Agency. Privacy and mobile device apps. https://www.cisa.gov/news-events/news/privacy-and-mobile-device-apps. Accessed: 2024-02-14.

[12] Anupam Das, Martin Degeling, Daniel Smullen, and Norman Sadeh. Personalized privacy assistants for the internet of things: Providing users with notice and choice. *IEEE Pervasive Computing*, 17(3):35–46, 2018.

[13] Pardis Emami-Naeini, Yuvraj Agarwal, Lorrie Faith Cranor, and Hanan Hibshi. Ask the experts: What should be on an IoT privacy and security label? In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 447–464. IEEE, 2020.

[14] Pardis Emami-Naeini, Janarth Dheenadhayalan, Yuvraj Agarwal, and Lorrie Faith Cranor. An informative security and privacy "nutrition" label for internet of things devices. *IEEE Security & Privacy*, 20(2):31–39, 2022.

[15] Jennifer Fereday and Eimear Muir-Cochrane. Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International Journal of Qualitative Methods*, 5(1):80–92, 2006.

[16] U.S. Food and Drug Administration (FDA). The nutrition facts label. https://www.fda.gov/food/nutrition-education-resources-materials/nutrition-facts-label. Accessed: 2024-04-21.

[17] Huiqing Fu, Yulong Yang, Nileema Shingte, Janne Lindqvist, and Marco Gruteser. A field study of run-time location access disclosures on android smartphones. In *Symposium on Usable Security and Privacy (USEC) 2023*, Feb 2014.

[18] Hamza Harkous, Kassem Fawaz, Remi Lebret, Florian Schaub, Kang G Shin, and Karl Aberer. Polisis: Automated analysis and presentation of privacy policies using deep learning. *arXiv preprint arXiv:1802.02561*, 2018.

[19] Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W Reeder. A "nutrition label" for privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, pages 1–12, 2009.

[20] Patrick Gage Kelley, Lucian Cesca, Joanna Bresee, and Lorrie Faith Cranor. Standardizing privacy notices: An online study of the nutrition label approach. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, page 1573–1582, New York, NY, USA, 2010. Association for Computing Machinery.

[21] Patrick Gage Kelley, Lorrie Faith Cranor, and Norman Sadeh. Privacy as part of the app decision-making process. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 3393–3402, 2013.

[22] MG Kendell. Rank correlation methods. *Charles Griffin and Company: London*, 1955.

[23] Rishabh Khandelwal, Asmit Nayak, Paul Chung, and Kassem Fawaz. Comparing privacy labels of applications in android and iOS. In *Proceedings of the 22nd Workshop on Privacy in the Electronic Society*, WPES '23, page 61–73, New York, NY, USA, 2023. Association for Computing Machinery.

[24] Kleimann Communication Group. Evolution of a Prototype Financial Privacy Notice: A Report on the Form Development Project, September 2010. [Online; posted 13-September-2012].

[25] Simon Koch, Malte Wessels, Benjamin Altpeter, Madita Olvermann, and Martin Johns. Keeping privacy labels honest. *Proc. Priv. Enhancing Technol.*, 2022(4):486–506, 2022.

[26] Konrad Kollnig, Anastasia Shuba, Max Van Kleek, Reuben Binns, and Nigel Shadbolt. Goodbye tracking? Impact of iOS app tracking transparency and privacy labels. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 508–520, New York, NY, USA, 2022. Association for Computing Machinery.

[27] Tianshi Li, Kayla Reiman, Yuvraj Agarwal, Lorrie Faith Cranor, and Jason I. Hong. Understanding challenges for developers to create accurate privacy nutrition labels. In *CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery.

[28] Jialiu Lin, Bin Liu, Norman Sadeh, and Jason I. Hong. Modeling users' mobile app privacy preferences: Restoring usability in a sea of permission settings. In *Proceedings of the Tenth Symposium on Usable Privacy and Security (SOUPS '14)*, pages 199–212, 2014.

[29] Yanzi Lin, Jaideep Juneja, Eleanor Birrell, and Lorrie Faith Cranor. Data safety vs. app privacy: Comparing the usability of android and ios privacy labels. In *Proc. Priv. Enhancing Technol.*, pages 182–210, 2024.

[30] Bin Liu, Mads Schaarup Andersen, Florian Schaub, Hazim Almuhimedi, Shikun (Aerin) Zhang, Norman Sadeh, Yuvraj Agarwal, and Alessandro Acquisti. Follow my recommendations: A personalized privacy assistant for mobile app permissions. In *Twelfth Symposium on Usable Privacy and Security (SOUPS '16)*, pages 27–41, 2016.

[31] Bin Liu, Jialiu Lin, and Norman Sadeh. Reconciling mobile app privacy and usability on smartphones: Could user privacy profiles help? In *Proceedings of the 23rd International Conference on World Wide Web (WWW '14)*, pages 201–212, New York, NY, USA, 2014.

[32] Aleecia M McDonald and Lorrie Faith Cranor. The cost of reading privacy policies. *I/S: A Journal of Law and Policy for the Information Society*, 4:543, 2008.

[33] J.H. McDonald and University of Delaware. *Handbook of Biological Statistics*. Sparky House Publishing, 2009.

[34] Ashwini Rao, Florian Schaub, Norman Sadeh, Alessandro Acquisti, and Ruogu Kang. Expecting the unexpected: Understanding mismatched privacy expectations online. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, pages 77–96, Denver, CO, June 2016. USENIX Association.

[35] Abhilasha Ravichander, Alan W Black, Thomas Norton, Shomir Wilson, and Norman Sadeh. Breaking down walls of text: How can NLP benefit consumer privacy? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4125–4140, 2021.

[36] Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. Question answering for privacy policies: Combining computational and legal perspectives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference*

on Natural Language Processing (EMNLP-IJCNLP), pages 4949–4959, Hong Kong, China, November 2019. Association for Computational Linguistics.

[37] Robert W. Reeder, Lujo Bauer, Lorrie F. Cranor, Michael K. Reiter, and Kami Vaniea. More than skin deep: Measuring effects of the underlying model on access-control system usability. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, page 2065–2074, New York, NY, USA, 2011. Association for Computing Machinery.

[38] Robert W. Reeder, Lujo Bauer, Lorrie Faith Cranor, Michael K. Reiter, Kelli Bacon, Keisha How, and Heather Strong. Expandable grids for visualizing and authoring computer security policies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, page 1473–1482, New York, NY, USA, 2008. Association for Computing Machinery.

[39] Robert W. Reeder, Patrick Gage Kelley, Aleecia M. McDonald, and Lorrie Faith Cranor. A user study of the expandable grid applied to P3P privacy policy visualization. In *Proceedings of the 7th ACM Workshop on Privacy in the Electronic Society*, WPES '08, page 45–54, New York, NY, USA, 2008. Association for Computing Machinery.

[40] Joel R Reidenberg, Travis Breaux, Lorrie Faith Cranor, Brian French, Amanda Grannis, James T Graves, Fei Liu, Aleecia McDonald, Thomas B Norton, and Rohan Ramanath. Disagreeable privacy policies: Mismatches between meaning and users' understanding. *Berkeley Tech. LJ*, 30:39, 2015.

[41] Daniel Smullen, Yuanyuan Feng, Shikun Zhang, and Norman M. Sadeh. The best of both worlds: Mitigating trade-offs between accuracy and user burden in capturing mobile app privacy preferences. *Proc. Priv. Enhancing Technol.*, 2020(1):195–215, 2020.

[42] Primal Wijesekera, Joel Reardon, Irwin Reyes, Lynn Tsai, Jung-Wei Chen, Nathan Good, David Wagner, Konstantin Beznosov, and Serge Egelman. Contextualizing privacy decisions for better prediction (and protection). In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*, pages 1–13, 2018.

[43] Shikun Zhang, Yuanyuan Feng, Anupam Das, Lujo Bauer, Lorrie Cranor, and Norman Sadeh. "Did you know this camera tracks your mood?": Understanding privacy expectations and preferences in the age of video analytics. *Proc. Priv. Enhancing Technol.*, 2021(2):282–304, 2021.

[44] Shikun Zhang, Yuanyuan Feng, Yaxing Yao, Lorrie Faith Cranor, and Norman Sadeh. How usable are iOS app privacy labels. *Proc. Priv. Enhancing Technol.*, 2022(4):204–228, 2022.

[45] Shikun Zhang and Norman Sadeh. Do privacy labels answer users' privacy questions? In *Symposium on Usable Security and Privacy (USEC) 2023*, Feb 2023.

A   Supplemental Tables and Figures

Gender		Age		Ethnicity	
Female	50.0%	18–25	21.5%	Asian	11.0%
Male	50.0%	26–35	34.0%	African American	7.0%
		36–45	23.5%	Caucasian	71.5%
		46–55	10.0%	Mixed	6.0%
		56–65	8.0%	Other	3.5%
		66+	3.0%	No data	1.0%

Table 3: Demographics of our study participants $N = 200$
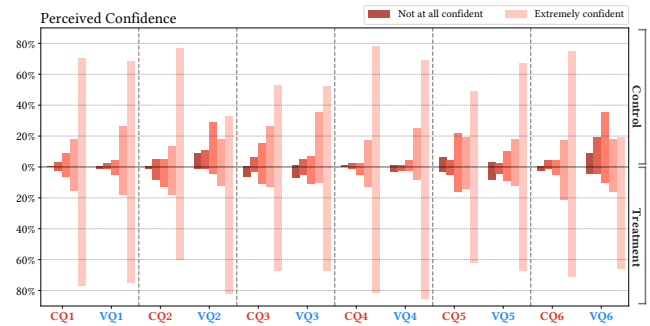
Figure 6: Distribution of participant confidence in their answers to 12 questions across both conditions. The top half represents the control group, while the bottom half shows the treatment group.

Question	Control %	Treatment %	Effect Size	Size per Condition
CQ1	0.35	0.74	0.82	26
VQ1	0	0.79	2.66	4
CQ2	1	0.74	0.82	26
VQ2	0.2	0.89	1.79	7
CQ3	0.65	0.63	0.04	10276
VQ3	0.55	0.68	0.27	227
CQ4	0.25	0.89	1.58	8
VQ4	0.45	0.95	1.26	12
CQ5	0.55	0.68	0.27	227
VQ5	0.7	0.74	0.08	2570
CQ6	1	0.6	1.10	15
VQ6	0.7	0.9	0.49	70

Table 4: A Priori Sample Size for 12 Questions Based on Pilot Results using G*Power

B   Survey Text

Consent Form

- I am at least 18 years of age.
- I have read and understand the consent information above.
- I want to participate in the research and continue with the survey.

Introduction   This survey is being conducted for research at Carnegie Mellon University. We will ask you to view two websites on your iPhone and answer questions about them. This survey should take about 20 minutes to complete. You will receive your compensation via Prolific upon completion of the study. To participate in this survey, you must use an iPhone with iOS 14 and above and have access to your iPhone throughout the duration of the survey. We recommend that you take this survey on a desktop, laptop, tablet, or other device besides your iPhone. Your participation is voluntary. Please do not reveal any private or personally-identifiable information about yourself or others during the survey.

- What is your Prolific ID? Please note that the text box should auto-fill with the correct Prolific ID.

General Questions about Privacy Label and Apps.   Please open the first link by scanning the QR code below with your iPhone's camera. If you are unable to scan the QR code, you cannot participate in this study and you will not get paid.

Please scroll down and view the App Privacy section of this page so that the privacy label is visible on your screen. We are going to ask you a few questions about this section, so please explore the label before continuing the survey.

- Have you seen an iOS app privacy label like this before?

- (Follow-up Yes) How often do you check privacy labels before downloading an app?

- Was privacy ever a reason you decided to not download or stop using an app?

To answer these questions, you will have to interact with the privacy label. Feel free to explore the label for as long as you would like before answering the following questions.

App Comprehension Questions

Q1: Does this app collect data for Analytics purposes and, if so, what data? (Select all that apply)

- Browsing History
- Contact Info
- Contacts

- Diagnostics
- Financial Info
- Health & Fitness
- Identifiers
- Location
- Other Data
- Purchases
- Search History
- Sensitive Info
- User Content
- Usage Data
- This app does not collect data used to track you or for tracking purposes
- I'm not sure

Q2: Does this app collect location data for Third-Party Ads/Advertising purposes?

- It collects precise location for Third-Party Ads purposes
- It collects coarse location for Third-Party Ads purposes
- It collects both precise and coarse location for Third-Party Ads purposes
- It does not collect location for Third-Party Ads purposes
- I'm not sure

Q3: Does this app collect Photo and Video data and, if so, for what purpose(s)? (Select all that apply)

- Analytics
- Developer Ads
- Functionality
- Other
- Personalization
- Third-Party Ads
- Tracking or Data Used to Track You
- This app does not collect [photo and video] data for any purpose
- I'm not sure

Q4: Does this app collect Purchases data and, if so, for what purpose(s)? (Select all that apply) [answers same as Q3]

Q5: Does this app link Diagnostics data to your identity?

- Yes
- No
- I'm not sure

Q6: Does this app collect Data Used to Track You or for Tracking purposes and, if so, what data? (Select all that apply) [answer choices same as Q1]

[For each of the 6 questions above, we asked the following 2 questions]

- What helped you to arrive at this answer? [short response]
- How confident do you feel that the answers you gave about the information on the privacy label are correct? Completely confident (5) to Not at all confident (1).

Second app prompt   We will now ask you to complete the same questions for a second app. You can access the second link by scanning the QR code below with your iPhone's camera. If you are unable to scan the QR code, you cannot participate in this study and you will not get paid. Please make sure that you scroll down to the App Privacy section so that the privacy label is visible.

[Repeat Q1 to Q6 for the 2nd app]

Treatment only Questions

QT1: Using the screenshots below, which app collects Diagnostics data and links it to your identity for any purpose? [Candy label only[4]] [Venmo label only[4]]

- App A collects Diagnostics data and links it to your identity
- App B collects Diagnostics data and links it to your identity
- Both apps collect Diagnostics data and link it to your identity
- Neither app collects Diagnostics data and links it to your identity
- I'm not sure

QT2: Using the screenshots below, which app collects more Usage Data for Analytics purposes?

[Candy label only[4]] [Venmo label only[4]]

- App A collects more Usage Data for Analytics purposes
- App B collects more Usage Data for Analytics purposes
- Both apps collect the same amount of Usage Data for Analytics purposes
- Neither app collects Usage Data for Analytics purposes
- I'm not sure

QT3: How useful were the colors in the grid as you answered the questions above?

- Very useful
- Moderately useful

---
[4]no legend

- Somewhat useful
- A little useful
- Not at all useful

QT4: Did you notice the legend? If so, did you use it?

- Yes I noticed it and used it as I answered the questions
- Yes I noticed it but did not use it to answer the questions
- I'm not sure if I saw it
- No I did not notice it
- Other [short response]

Term Definition Questions

QTD1: What does "data linked to you" mean?

- Data that is transferred when you use an app and stored in a database
- Data from your account or device that could be used to identify you
- Data that is used to track you and your activity while using the app
- Information you've given during the sign-up process of an app
- Data that includes your real name, or phone number, or address
- I'm not sure
- Other [short response]

QTD2: What does "data not linked to you" mean?

- Data that is not personal information, but could be used to determine information about you
- Contact information, such as an email address or phone number
- Data not connected to you, even if it is collected by the app
- Data that developers can use to identify you, but is not shared with third parties
- Data that does not include your real name or location
- I'm not sure
- Other [short response]

QTD3: What does "data used to track you" mean?

- Identifiable data that is shared with third parties to personalize ads
- Your location and physical address are collected by the app
- Patterns of using an app, such as frequency or search history

- Data sent to third parties only for security purposes
- I'm not sure
- Other [short response]

QTD4: If an app shared your location and email address with third party advertisers, do you think that would be considered "tracking"? Yes/No/I'm not sure

QTD5: If an app used your location to show you nearby stores, do you think that would be considered "tracking"? Yes/No/I'm not sure

General Perceptions

QGP1: How helpful did you find the privacy label to be?

QGP2: Generally, how easy or difficult was it to understand the privacy labels? Very Easy (1) to Very Difficult (5)

QGP3: Please rate how easy each element of the privacy label is to understand on a scale from Very Easy (1) to Very Difficult (5). [matrix question]

- Terms used in the label
- Finding definitions of terms used in the label
- Icons [control] / Colors [treatment]
- Label structure

QGP4: Was any part of the label confusing, and if so, please explain. [short answer]

QGP5: Do you think this privacy label provides enough information about how an app collects and uses your data? Yes; No; I'm not sure; Other [Follow up if No] What information would you like to see added to the label, if any? [short answer]

QGP6: If you have any suggestions for improving the privacy label, please provide them below.

QGP7: In the future, do you plan to look at these labels before deciding to download an app?

QGP8: Do you have any other comments or feedback regarding the privacy labels or the survey? [short answer]

Wrap-up   You will receive payment on Prolific for completing this survey. We thank you for your time spent taking this survey. Your response has been recorded.

C   Study Screenshots

(a) Control Venmo

(b) Control Candy

(c) Treatment Venmo

(d) Treatment Candy

Figure 7: Webpages shown to participants

Privacy Requirements and Realities of Digital Public Goods

Geetika Gopi
Carnegie Mellon University

Aadyaa Maddi
Carnegie Mellon University

Omkhar Arasaratnam
OpenSSF

Giulia Fanti
Carnegie Mellon University

Abstract

In the international development community, the term "digital public goods" is used to describe open-source digital products (e.g., software, datasets) that aim to address the United Nations (UN) Sustainable Development Goals. DPGs are increasingly being used to deliver government services around the world (e.g., ID management, healthcare registration). Because DPGs may handle sensitive data, the UN has established user privacy as a first-order requirement for DPGs. The privacy risks of DPGs are currently managed in part by the DPG standard, which includes a prerequisite questionnaire with questions designed to evaluate a DPG's privacy posture.

This study examines the effectiveness of the current DPG standard for ensuring adequate privacy protections. We present a systematic assessment of responses from DPGs regarding their protections of users' privacy. We also present in-depth case studies from three widely-used DPGs to identify privacy threats and compare this to their responses to the DPG standard. Our findings reveal serious limitations in the current DPG standard's evaluation approach. We conclude by presenting preliminary recommendations and suggestions for strengthening the DPG standard as it relates to privacy. Additionally, we hope this study encourages more usable privacy research on communicating privacy, not only to end users but also third-party adopters of user-facing technologies.

1 Introduction

Today, digital government services—like national registries, payment systems, or healthcare systems—are often imple-

mented and administered by third-party vendors [50]. This comes with a few drawbacks. Vendors are known to charge high prices [9], and governments are subsequently subject to vendor lock-in, due either to monopolies or a lack of interoperability between market offerings [1]. These costs are typically passed on to residents in the form of taxation, which can be particularly problematic in low-income countries [29].

Digital Public Goods (DPGs) are a concept that was recently revived in the international development community, partially to counter this trend. DPGs are open-source digital goods that are designed to benefit society [19]. In 2020, the United Nations (UN) put forth a report calling for "a platform for sharing digital public goods... in a manner that respects privacy, in areas related to attaining the Sustainable Development Goals." [33]. In response, the Digital Public Goods Alliance (DPGA) was formed to encourage and steward the development of DPGs. Precisely, the DPGA defines DPGs as "open source software, open data, open AI models, open standards and open content that adhere to privacy and other applicable laws and best practices, do no harm, and help attain the [United Nations' Sustainable Development Goals]" [16].

In the years since the DPGA was formed, DPGs have occupied a growing role in government services worldwide, as well as other community-driven services. For example, MOSIP is a DPG digital ID system with over a 100 million users, currently adopted by 11 countries [52]. DIGIT HCM is a health campaign management DPG with over 15 million users [24]. DIVOC is another health campaign management DPG with over 160 million users and is currently adopted by 5 countries including India, Philippines and Sri Lanka [24].

Privacy is a first-order concern in DPGs. In addition to being highlighted as a key property in the original UN report [33], it is also a central component of the evaluation used to select which projects are officially listed as DPGs [6]. Briefly, the evaluation proceeds as follows (detailed description in Section 2): a DPG candidate project first submits answers to an official DPG questionnaire, which contains 24 questions [6]. One of these questions specifically asks about what personally identifiable information (PII) the DPG

candidate collects, and how this information is protected. Responses from the candidate are meant to be backed up by supporting documentation and/or code. DPG candidates' responses to the questionnaire are then reviewed by the DPGA; if the responses are judged to be sufficiently high-quality and consistent with the DPGA's requirements, the candidate is formally approved as a DPG. Although DPG certification does not convey any explicit rights or privileges to the holders, it appears to be used in practice as a form of advertisement, e.g., by being listed on the DPG's website [24, 35, 52].

In this paper, we study whether the DPG approval process is effective at selecting DPGs that protect **user privacy**. We evaluate this in two phases: first, we run a qualitative study to analyze the responses of DPGs to the privacy component of the questionnaire. We evaluate these responses in terms of their completeness and their adherence to established privacy best practices. We then conduct an in-depth case study on three DPGs, in which we analyze their structure, documentation, and possible privacy threats. We use this analysis to determine whether DPGs may have privacy implications that are not captured by their responses to the DPG standard.

Our results show that existing DPGs provide a wide range of responses to the questionnaire, many of which convey limited maturity or attention to privacy and data protection. This suggests that the current DPG standard, and the associated approval process, does not filter out projects that take a lax approach to privacy. We emphasize that the DPGA is in a difficult position: it is neither a standards agency nor an enforcement agency. It is unrealistic to expect that it will be able to evaluate the privacy properties of candidate DPGs, many of which comprise complex code bases and documentation. Nonetheless, the UN has outlined privacy as a first-order requirement for DPGs. Hence, we believe it is important to find a solution that both encourages privacy best practices from DPGs, while also working within the existing constraints.

Contributions Our contributions in this paper are threefold:

1. We conduct a qualitative study of the privacy responses of 101 DPGs. We code their responses according to (a) high-level qualitative properties, and (b) common privacy themes that were extracted from existing privacy frameworks. We find that a high percentage (40% of DPGs) did not provide an adequate level of detail to understand how they handle PII. Moreover, many DPGs make common mistakes, such as conflating privacy compliance with privacy protection (50%) and shifting responsibility for data stewardship to other parties (17%).

2. We conduct three in-depth case studies of DPGs with over 1 million users and from different sectors. Among these, we find that even mature DPGs that answered the DPG standard thoroughly can have gaps in how their documentation addresses privacy concerns. These gaps have been communicated to the relevant DPGs.

3. We make several recommendations for how to improve the DPG standard to encourage better privacy protections. At a high level, our main recommendation consists of requiring a more detailed privacy assessment (akin to Privacy Impact Assessments [64]), to be completed by third parties or the DPGs themselves. The DPGA would no longer evaluate the quality of privacy responses, but would provide the privacy assessment documentation on their website for downstream users to evaluate. These recommendations (outlined in more detail in Section 6.2) have been communicated with the DPGA, and are currently under consideration for a restructuring of the DPG standard.

Usable privacy research often focuses on privacy for end-users. However, for DPGs, there are several stakeholders that want to (a) demonstrate that their methods are private (DPG candidates) and (b) evaluate the privacy claims of other organizations (DPGA) who also need those processes to be "usable". While existing research has studied how to communicate privacy to end users (e.g., through privacy nutrition labels [43, 49]), little research has been conducted on communicating privacy to third-party adopters, who have more technical sophistication than a typical user, but less sophistication than a domain expert. This area is relatively under-explored, and the DPG environment, being open-source, is an excellent opportunity to study such questions.

2 Background and Related Work

2.1 The DPG Standard and Questionnaire for Privacy

Assessing and endorsing a DPG candidate involves a three-step process. The organization seeking DPG status must first complete an online application on the DPGA's website, including a DPG questionnaire [17]. As part of the application, candidates are required to submit various forms of supporting evidence such as technical documentation, open licenses, and privacy policies. Once the application is received, it undergoes an evaluation process based on the DPG standard [6], which serves as a set of specifications and guidelines that defines a Digital Public Good. To receive recognition from the DPGA and the wider community, a DPG candidate must meet the baseline requirements as outlined in the DPG standard. If the application satisfies all the criteria of the DPG standard, it will be acknowledged as a Digital Public Good and included in the DPG registry [5].

The evaluation process for the DPG standard involves a thorough assessment of various criteria including accessibility, functionality, interoperability, and privacy, among others. The DPG standard is an open-source standard maintained by the DPGA [6]. Its credibility is further endorsed by a growing list of experts who advocate for open-source entities [6].

The DPG standard [6] includes three privacy-related sections, i.e., Sections 7, 8 and 9, outlining the privacy requirements for a DPG candidate. Section 9(a) specifically addresses data privacy and security by requiring DPGs to demonstrate how they ensure the privacy, security and integrity of personal information collected, stored and distributed as part of their solution. Section 7 asks candidates to explain how they ensure compliance with relevant privacy and applicable laws. Sections 8 and 9 of the DPG standard further require candidates to explain their efforts to follow best practices and ensure that the solution does no harm to their users.

The DPG standard is implemented via the DPG questionnaire [6], which evaluates candidates' adherence to the 9 indicators of the DPG standard. The questionnaire comprises both open-ended and multiple-choice questions. In this study, we are most interested in Section 9(a), which requires candidates to respond to the following question:

> "How does your solution ensure data privacy & security? Please demonstrate how the project ensures the privacy, security and integrity of this data and the steps taken to prevent adverse impacts resulting from its collection, storage and distribution." (open form)

This open-ended question leaves much room for interpretation, and does not precisely define what is meant by privacy. Hence, our primary research question for this paper is as follows:

> *Does the current DPG standard effectively evaluate or document digital solutions' potential privacy harms?*

To study this question further, we first conducted a qualitative analysis of 101 DPG responses to the privacy component of the questionnaire. This is described further in Section 4. We then conducted an in-depth case study on three DPGs to elicit possible privacy threats that the DPG questionnaire fails to capture. This is described further in Section 5.

2.2 Related Work

Evaluating the Potential and Drawbacks of DPGs While DPGs have not received as much attention from the academic research community, several papers (many of them position papers) highlight the significance of DPGs and the factors that impact their utility [11, 21, 34, 45, 59, 62, 63, 74, 77, 79]. Nickholson *et al.* explored key challenges and opportunities in achieving the Sustainable Development Goals through DPGs [62]. Their writing emphasizes that the potential harm from a DPG is not only associated with the technology itself, but also depends on its implementation, usage, and evolution over time, highlighting the need for further research in the DPG space [62]. This observation closely aligns with our

own findings, and partially motivates this study. Mukherjee *et al.* describe case studies that illustrate the concept of digital building blocks as public goods and demonstrate their application to developmental challenges such as poverty, inequality, health, education, public administration, and governance that affect entire populations [59]. Kumar et al. and Chen et al. explore the factors influencing contributions to DPGs, while also highlighting the importance of enhancing the quality of DPGs [11, 45]. These studies conducted large-scale field experiments and employed power analysis methods to study the correlation between factors influencing experts' contributions to DPGs.

Incentivizing Privacy Best Practices While there has been a vast literature studying how to incentivize organizations to invest in cybersecurity [30, 44, 48, 73], there has been comparatively less work analyzing economic incentives of organizations to invest in data privacy [4, 47, 91]. Instead, many organizations' policies and procedures surrounding data privacy are primarily driven by compliance with privacy regulation, either directly [3, 37, 76, 81, 89, 91] or indirectly, e.g., via vendor requirements [7, 64]. However, compliance with privacy regulations *does not inherently ensure that an organization is adequately protecting user privacy* [38]. While there is not a single global standard for data privacy, many existing privacy frameworks (e.g., LINDDUN [92]) present compliance as only one part of a robust privacy posture [64, 65]. Indeed, empirical observations show how various components of a privacy strategy, beyond just compliance, can interact to affect the utility of a product. For instance, Adjerid *et al.* showed that privacy regulation combined with collecting proper consent from users can actually result in greater data sharing than under fully unregulated situations [4]. This suggests the importance of coupling structured privacy requirements around privacy with a clear mechanism for collecting user consent. At the same time, the very act of asking for consent can affect users' willingess to share their data with a service, as demonstrated by Lam *et al.* with regards to the opt-in requirement of GDPR [47].

While our recommendations in Section 6.2 relate to incentivizing privacy best practices, this paper focuses on the higher-level question of whether the current DPGA standard ensures that DPGs' privacy postures are consistent with the recommendations of prominent privacy frameworks.

3 Methodology

Our evaluation is split into two components.

1. **Qualitative Analysis of DPG Responses (§4)** We first conducted a qualitative document analysis of DPG responses from all approved DPGs as of May 12, 2023. Our goal was to understand trends in the content and quality of DPG candidates' responses to the privacy question.

2. **DPG Case Studies (§5)** We next conducted detailed case studies into three DPGs to understand their privacy implications. The case studies were conducted on August, November, and December 2023, respectively. Our goal for this component was to understand how responses in the DPG standard are correlated (or not) with actual implementations or architectures.

We next detail the methodology for each component. We discuss the limitations of our methodology in Section 6.1.

3.1 Methods: Qualitative Analysis of DPG Responses

To analyze DPG responses, we first gathered all 167 DPG responses from the DPGA's GitHub repository on May 12, 2023 [18] and filtered them to include only DPGs indicating the collection of personally identifiable information (PII), as these are the only ones that answer Section 9(a). This resulted in a total of 101 relevant DPG responses. Filtering was needed because DPG candidates that did not indicate collection of PII would have no further statements to analyze regarding privacy. We could not access rejected DPG responses.

Analysis The lead researcher coded 50 responses independently to develop the initial codebook. The lead researcher developed 'Privacy Component Analysis' codes with a priori coding, using existing privacy frameworks (e.g., LINDDUN, APEC) [13, 92]. For remaining themes, the lead researcher used emergent coding. The lead and second researcher went over the coded responses and refined the codebook through discussions. The two coders coded the remaining 51 responses independently using the final codebook (Appendix A). The inter-rater reliability (IRR) was computed over these responses using percentage agreement (responses coded the same way, divided by the total number of responses). We achieved an inter-rater reliability of 0.87, which is considered acceptable [68]. Since codes were not mutually exclusive, Cohen's kappa was inapplicable. Conflicts were resolved through discussion. We emphasize that our study size is relatively small (101 DPGs), but consists of the *entire* population of DPGs that claimed to collect PII at the time of data collection. Hence, we present counts of occurrences of codes.

3.2 Methods: Detailed Case Studies

For our case studies, we chose three DPGs based on specific criteria: (1) having a user base of over 1 million users, and (2) providing documentation with specific sections related to privacy. Since these large-scale systems are labor-intensive to review and analyze, we decided to focus on 3 DPGs with significant impact in user-facing sectors: healthcare, digital IDs, and news and media.

Analysis Our case studies used privacy threat modeling [92], a structured approach used to identify potential privacy threats

within a system or application. This involves analyzing the system's components, data flows, and potential vulnerabilities that could compromise user privacy. Using these techniques, we compared our findings with the responses provided by DPGs to the DPG standard. Our goal is to understand how well responses to the questionnaire relate to a more detailed analysis of the DPG. Below, we outline our three-step methodology for identifying privacy threats.

We first reviewed the technical documentation from the selected DPGs and identified all system components involved in processing personal data. This step gave us a thorough understanding of the DPG's system architecture. Next, we carefully identified and analyzed the data flows [41] by creating Level 2 data flow diagrams [32]. Using Level 2 diagrams lets us capture potential privacy risks without considering low-level system details. Finally, we used the LINDDUN threat modeling framework [92] to identify potential privacy threats. We highlight that this methodology is capturing potential privacy vulnerabilities that are implied by the documentation. It does *not* necessarily imply that a vulnerability actually exists in the software. For example, some of the vulnerabilities we found were confirmed to be documentation mistakes (not true vulnerabilities) by the DPGs.

The case study results provided valuable insights into the actual privacy practices and strategies employed by the selected DPGs, shedding light on the effectiveness of their privacy protection mechanisms. We compared this analysis to DPGs' responses on the questionnaire to explore whether DPGs can have privacy implications that are not captured by the DPG standard.

4   Qualitative Analysis of DPG Responses

When analyzing the responses of DPG candidates to Section 9(a) of the DPG questionnaire, we generated codes related to four main themes (codebook construction in §3.1):

- **Overall response quality.** Were the responses clear, internally consistent, and specific?

- **Types of supporting documentation.** What kind of supporting documentation did the DPG candidate provide?

- **Proposed privacy safeguards.** What technical and process strategies were used by DPGs to protect user data?

- **Coverage of privacy best practices.** Did the response cover common elements of existing privacy frameworks and principles?

These themes were identified using a top-down approach to answer two questions: (1) How did the candidates respond — both in their main response ("Proposed privacy safeguards") and "Types of supporting documentation", and 2) How well did they respond, in form ("Overall response quality") and function ("Coverage of privacy best practices").

These themes helped us understand how DPGs approach privacy and whether the current evaluation process helps the
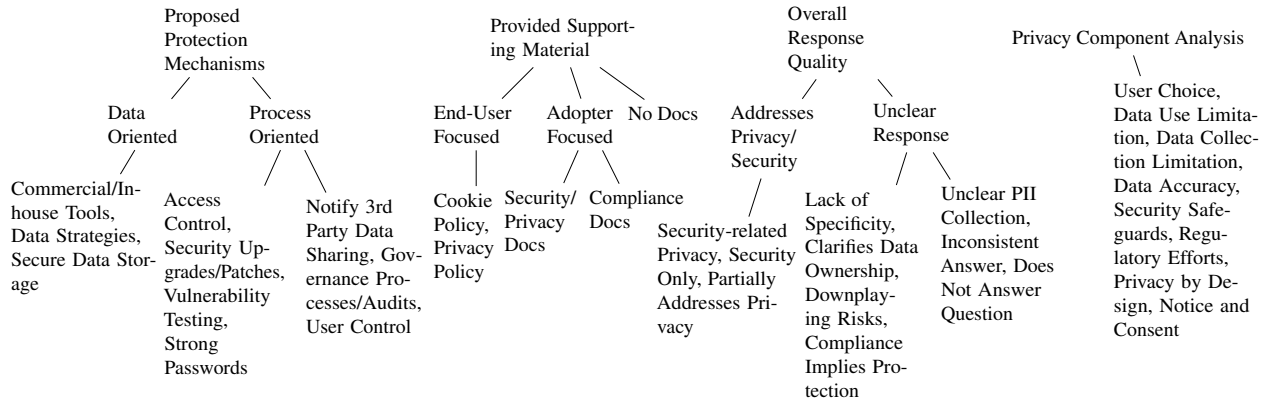
Figure 1: Categorizing codes under the four themes we consider during qualitative analysis of DPG responses.
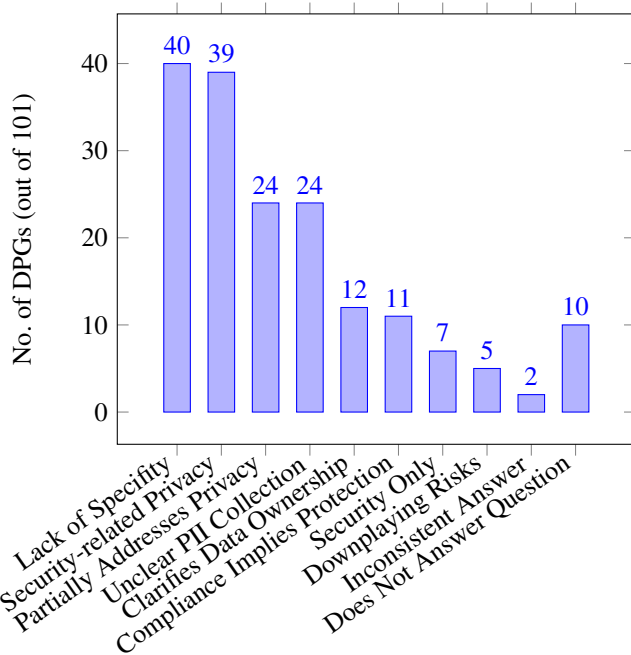
Figure 2: Results from qualitative analysis of DPG responses for Overall Response Quality.

DPGA screen out DPGs with possible privacy threats. A categorization of our identified codes within the four themes is illustrated in Figure 1. We next present our results, divided according to theme.

4.1 Overall Response Quality

When evaluating the overall quality of responses, we observed that the majority were either vague, solely focused on security controls, or only partially addressed privacy controls. Roughly, we categorized the responses as ones that 'address security/privacy' or were 'unclear responses'. Our results for overall response quality are illustrated in Figure 2.

Notably, we found that 40% of DPGs lacked specificity in describing their protection methods. We defined 'lack of specificity' as responses that mentioned privacy-related terms such as 'anonymization' or 'obfuscation', without explaining how it is applied within the context of the DPG solution. This could be attributed to the open-ended nature of the questionnaire. Refer to the codebook in Table 3 for the full list of response types and their definitions. An example of a response coded as lacking specificity is as follows:

> **DPG15**: "The solution promotes best security and quality assurance practices in an effort to support the privacy of PII and prevent adverse impact related to PII. Security and quality assurance best practices that can contribute to the prevention of adverse impact related to PII are integrated into our development processes and automated as possible."

Furthermore, 50% of DPGs appeared to primarily emphasize security-related privacy controls such as encryption, hashing, or regulatory measures as their main privacy strategies:

> **DPG43**: "All data transfer through HTTPS (SSL) & user level security is maintained through SHA-512 encryption with roles & privileges."

> **DPG80**: "While we don't collect the data ourselves, the software has high degrees of security and compliance at the software and network level to ensure data integrity."

While security measures and best practices are useful, they are not sufficient for guaranteeing data privacy. Interestingly, there was little mention of data-oriented strategies like data minimization and anonymization.

Among the stronger responses, about 25% of responses took steps that fully or partially addressed privacy by design principles. For instance:

DPG81: "We are collecting anonymized data (clinical data of patients) with prior approvals and clearances from hospitals."

On the other hand, we encountered several responses that showed a lack of understanding of privacy by design. 17% of DPGs appear to shift the responsibility for privacy to solution implementers; others downplayed privacy risk:

DPG101: "... Unfortunately, there is no such thing as true data protection, even when data is locally stored and hosted in a country..."

DPG11: "As a default, this project does not collect or store PII data, but some partners and deployments would like the option to have the same; in which case we store the name, address and phone number of the consenting individual and clearly mention in our contracting terms that we do not own any of this data"

Overall, these responses reveal a wide spectrum of qualities in responses. Most importantly, they suggest that **many DPGs are currently not providing enough detail for the DPGA or an adopter to understand its privacy posture.**

4.2 Provided Supporting Material

We found that few DPGs provided supporting documentation, and those that did often provided policies. We categorize these as 'end-user focused' or 'adopter focused' material. End-user focused documentation refers to material that is seen by individuals whose PII can be handled by the DPG. Adopter focused documentation, on the other hand, can be technical material informing DPG implementers how to use the supported security and privacy measures, or contain instructions for compliance with privacy regulations such as GDPR.

As shown in Figure 3a, over 50% of the studied DPGs did not provide any supporting documentation to explain their protection mechanisms, and only about 16% of DPGs submitted some form of documentation related to security or privacy. The remaining DPGs submitted privacy policies, cookie policies, or compliance-related documentation. These results suggest that some DPGs may equate privacy compliance with privacy protection. They focused more on demonstrating compliance with regulations, rather than implementing robust privacy protection measures—a common phenomenon in security and privacy compliance [14, 90]. Refer to the codebook in Table 2 for the full list of supporting documentation types and their definitions.

4.3 Proposed Protection Mechanisms

We next turn to the tools and methods within DPG candidates' responses. Roughly, the privacy protection strategies

they proposed can be categorized as 'data-oriented' strategies (e.g., data strategies, secure data storage) and 'process-oriented' strategies (e.g., governance processes/audits, vulnerability testing). Data-oriented strategies are technical privacy measures that directly operate on data [40]. Process-oriented strategies, on the other hand, are organizational procedures that ensure responsible handling of data [40]. The full list of privacy protection mechanisms and their definitions is provided in Table 1.

As shown in Figure 3b, the most common privacy protection mechanisms claimed by DPGs were data strategies, secure data storage, and access control – around 26% of the DPGs mention using mechanisms that fall under one or more of these categories. DPGs that use data strategies mention techniques such as minimizing the amount of data collected and anonymizing any personal data collected. Secure data storage mechanisms involve using measures like encryption to protect personal data. Mechanisms under access control enforce restrictions for accessing personal data based on predefined rules and policies.

Nearly 9% of the DPGs propose taking responsibility for establishing and/or adhering to a governance process to ensure the protection of personal data. About 6% of DPGs proposed to implement user controls to let users express privacy preferences effectively (e.g., provide consent, submit data deletion requests). Less common strategies (1 - 3%) include routinely applying security upgrades and patches, and performing vulnerability testing to ensure user data is protected.

4.4 Privacy Component Analysis

Our final theme conducted a privacy coverage analysis, which was meant to understand whether DPGs are addressing common privacy considerations that arise in existing evaluation frameworks and guidelines. Since there is no single globally-adopted privacy framework or guideline, we extracted common components from five widely-used privacy frameworks and principles: the NIST Privacy Framework [65], LIND-DUN [92], APEC Information Privacy Principles [13], Privacy By Design Principles [10, 40], and principles outlined by the GDPR under Article 5 [42]. We assigned one code to each concept or idea that appears in *all* of the above resources, resulting in eight common components, listed in Figure 1. The definitions of these common components are provided in Table 4 (Appendix A).

In our analysis, we checked whether the DPGs addressed each of these components. We define 'addressing a component' as including some amount of documentation describing their efforts related to that component. Results from our analysis indicate that most of the components do not achieve a high coverage rate.

We find that **coverage of these common privacy components is sparse at best**. The component with the highest coverage rate across DPGs was security safeguards (55%), fol-
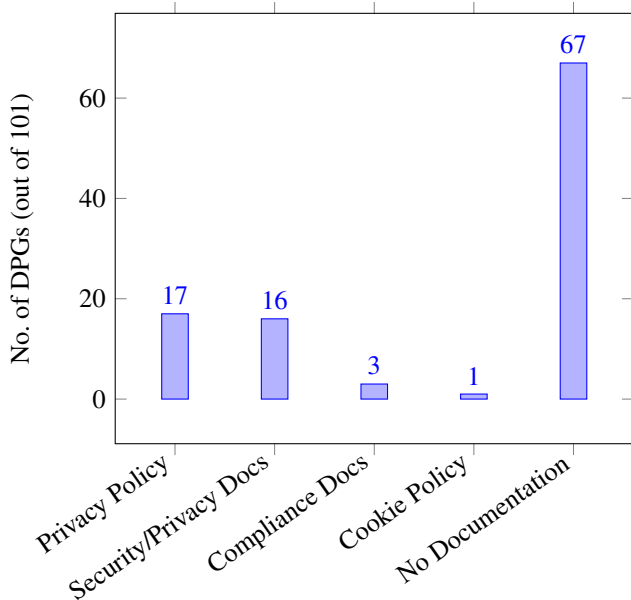
(a) Provided Supporting Material
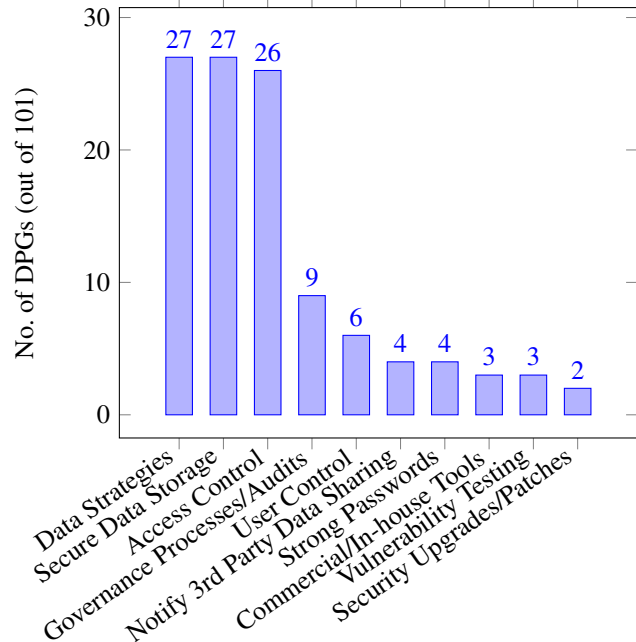
(b) Proposed Protection Mechanisms

Figure 3: Results from qualitative analysis of DPG responses.

lowed by regulatory efforts (33%). Most DPGs that addressed the security safeguards component employed measures like access control and encryption to protect user information. DPGs that addressed the regulatory efforts component mentioned their efforts to comply with regulations like GDPR, and some DPGs mentioned the use of audits to review their compliance to these regulations. These results reinforce our earlier observation that some DPGs equate privacy protection with security safeguards and compliance efforts. For example:

> **DPG18**: "... prioritizes security, stability, and scalability above all else, and many of our users implement ... to comply with GDPR, HIPAA, and other policies."

> **DPG21**: "We ensure the security of collected data both by technical and policy means ... the number of personnel who have access to such data is strictly limited ... all personnel have signed non-disclosure agreements (NDAs), which clearly define user's personal data as confidential information subject to confidentiality terms. The NDAs also imply strict monetary penalties in case of a breach. As to the technical level of ensuring security of the data, we use SSL certificates, database connections are private -> connection to the DB is available only via local server (outbound connections are disabled)."

We also find that DPGs largely did not on address user notice, choice, and control. Only 17% of the DPGs included
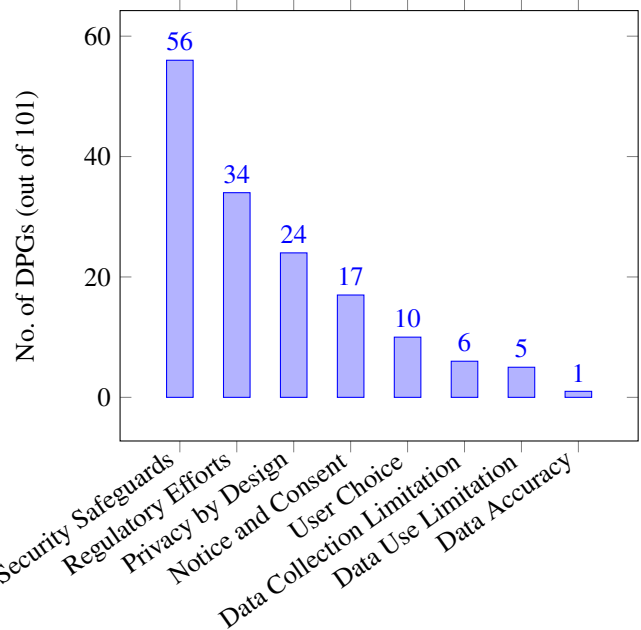
Figure 4: Results from qualitative analysis of DPG responses for Privacy Component Analysis.

documentation on notice and consent. Most DPGs that addressed this component mentioned that information on their data practices, such as purpose of data collection or third-party sharing, could be found in their privacy policies.

Furthermore, only around 10% of the DPGs addressed user choice in their responses to the questionnaire. These DPGs provided their users with some amount of control in how their data is collected, used, or shared. For instance:

> **DPG54**: "The demographic data (birth year and gender) fields are optional, and are not prerequisites for using the platform, allowing users for whom this information is more sensitive to opt out."

> **DPG75**: "All information is transferred securely using HTTPS and raw data provided by the user for analysis can be deleted at the user's request."

Few DPGs allowed users to delete their data upon request.

5 Case Studies of Digital Public Goods

To gain a deeper understanding of the actual privacy practices and strategies employed by DPGs, we conducted an in-depth case study of three different DPGs. The case-studies allowed us to determine whether DPGs may have possible privacy implications that are not captured by the DPG standard. **We re-emphasize that our findings are based only on documentation, and do not necessarily mean that the implementations have privacy vulnerabilities.**

5.1 Case Study 1: MOSIP

The Modular Open Source Identity Platform (MOSIP) is an open-source and open standards foundational identity platform [52]. It serves as an API-first platform for governments to build their own national ID platforms, offering ID life-cycle management and identity verification capabilities. The platform has over 100 million registered users and is operational in 11 countries, including Morocco, Ethiopia, and Sri Lanka.

Summary of Findings MOSIP's response to Section 9(a)(iii) of the DPG questionnaire states that "privacy and security practices are central to MOSIP and the project has taken extensive measures to provide security of data and has numerous existing and evolving features on privacy and data protection." MOSIP's response includes a link to its adopter-focused security and privacy documentation, which outlines the access control (e.g., authentication, rate-limiting) and secure data storage (e.g., encryption at rest) measures it supports.

The MOSIP response and documentation was among the more careful of the DPGs we analyzed. At the same time, our threat elicitation process revealed potential issues, such as data being revealed to third parties in plaintext during authentication and secure storage not being used at all stages of data ingestion. MOSIP reported that most of our findings were mistakes in the documentation (not in the underlying software), some of which have since been updated. Nonetheless, the information collected by the DPG standard is not nuanced enough to reveal such potential privacy vulnerabilities.

5.1.1 Analysis

The high-level architecture of MOSIP consists of two core modules: (1) ID Lifecycle Management, and (2) Authentication. ID lifecycle management includes several subcomponents such as ID pre-registration, enrollment, updation and de-activation [57]. The authentication module provides ID authentication services [56]. The data flow diagram (DFD) for MOSIP is illustrated in Figure 6 in the Appendix. This illustration informed the threat elicitation process using the LINDDUN framework [92].

Residents have the option to pre-register online and then visit designated centers to complete the registration process [55]. According to MOSIP's responses to the DPG questionnaire, the ID creation process needs residents to submit their legal name, age, address, biometrics (e.g., fingerprint, face, iris), and other PII as required by the country. When residents need to authenticate themselves with relying parties, these institutions serve as proxies to verify the residents digital IDs against MOSIP's servers [56].

Observation 1: Passing Clear Text Credentials to Relying Parties Authenticating a digital ID in MOSIP involves a relying party acting as a proxy to transmit credentials on behalf of the end-user. The relying party collects *unencrypted* end-user virtual IDs (VIDs) and one-time passwords (OTPs) and submits them to MOSIP's servers for verification. For this purpose, MOSIP utilizes a "yes/no" API to deliver verification responses and places trust in relying parties that may belong to private or government organizations [56].

However, over-reliance on these parties can lead to the misuse of user credentials received in clear text, allowing them to identify users even when temporary VIDs are used. This poses a privacy risk as the clear text credentials could be intercepted, compromising the identity and personal information of users as noted in prior work, which found a related vulnerability in OAuth 2.0 implementations [75]. Note that MOSIP offers an alternative authentication mechanism called eSignet, which mitigates this risk.

Observation 2: Weak Anonymization in Profiling System MOSIP offers an 'Anonymous Profiling System' [54] for conducting privacy-preserving analytics on pre-registration data. The anonymized dataset [54] includes personal information attributes like gender, location, and year of birth. Documentation indicates that anonymization is provided through suppression of data. Suppression is a form of weak anonymization that could introduce potential privacy vulnerabilities, as malicious actors may carry out reconstruction attacks [20] by launching targeted queries against the profiling system.

MOSIP counters the risk of linkage attacks by encrypting the database so that a record is unidentifiable without knowledge of the corresponding VID. Nonetheless, depending on which fields are shared with third parties, inference attacks using correlated data sources have been used in other contexts to de-anonymize users based on partial information [12, 60, 78], as well as inferring properties of groups of users.

Observation 3: Unencrypted Storage of Pre-Registration Data MOSIP's 'pre-registration' databases are downloaded to an operator's system for offline data retrieval [54]. However, in August 2023, at the time this case study was conducted, the documentation suggested that these databases are stored in an unencrypted format, without providing a justification for doing so [54]. The documentation has been updated since we shared our findings with MOSIP in September 2023; as of June 2024, it states that pre-registration data is indeed stored in encrypted form.

Observation 4: Unclear Documentation of Data Retention and Deletion Policies The documentation on MOSIP's data retention and deletion policies is unclear, as it uses two different terms: 'deactivation' and 'decommission'. Deactivation refers to temporary shutdown, while decommission refers to permanent shutdown of a resource [53]. It is unclear which option (if any) leads to permanent deletion of user data, including biometrics.

Observation 5: Possibly Low-Quality Informed Consent During registration, the operator can choose to mark consent on behalf of the individual [58]. This raises concerns about the quality of informed consent [31], as operators could mark consent without clearly explaining the terms to individuals.

Responsible Disclosure We communicated our observations with MOSIP, who confirmed that Observation 3 was a documentation gap. MOSIP has since updated that documentation, and more generally, significantly clarified their documentation of privacy data flows compared to when we ran this study.

5.2 Case Study 2: Ushahidi

Ushahidi is a crowd-sourcing platform for social activism. It aims to map and document information during political campaigns, natural disasters, and other events of public interest [82]. The platform enables local observers to easily submit reports via their mobile phones or the internet, creating an archive of events accompanied by geographic and time-date details. Ushahidi has been deployed in over 60 countries and supports more than 40 languages Some of its use cases include supporting earthquake relief efforts in Nepal, ensuring fair elections in Nigeria, and helping women address sexual violence in Egypt [85].

Summary of Findings Ushahidi's response to Section 9(a)(iii) of the DPG questionnaire consists of links to their documentation on how their platform supports data security

and the measures implementers must take to comply with GDPR. Specifically, their response describes "reasonable administrative, physical and electronic measures" like encrypting data in transit, securing servers using access control mechanisms like (i) restricting open ports, (ii) using hardened SSL configurations, and limiting communication between services to internal private networks. Ushahidi also provides implementers with instructions on collecting consent from users.

However, their response does not provide details about privacy measures (e.g., anonymization) for PII collected from sources other than surveys (e.g., Twitter, emails). Of concern, this data may still be stored on the platform even after the original data sources have been deleted. Moreover, their response states they use security safeguards during transit, but whether they encrypt this data at rest is unclear.

5.2.1 Analysis

Ushahidi's high-level architecture consists of three core components: the Platform, Services, and Data. The 'Platform' component includes Ushahidi's core platform and MySQL data store [86]. The 'Services' component provides POST and REST APIs for ingesting data and transmitting reports to Ushahidi's web interface and mobile application app [86]. According to Ushahidi's responses to the DPG questionnaire, the platform can collect "email addresses, location, and telephone numbers" of its users.

The 'Data' component allows implementers to configure input data sources [86]. End-users can submit reports via Ushahidi's web interface or send reports to dedicated email or SMS channels. Additionally, the platform can be integrated with Twitter (now known as X) to ingest data based on hashtags [88]. The data flow diagram (DFD) for Ushahidi is illustrated in Figure 7.

Observation 1: Inconsistent Data Updates Ushahidi supports the use of Twitter's (now known as X) developer API to collect messages (or tweets) based on hashtags [88]. This functionality aids in monitoring crisis response, elections, political and community engagements. The content from collected tweets is stored in a database called 'messages'. It is observed that content from deleted or modified tweets does not get updated on Ushahidi's platform [88]. As a result, data stored on the platform may become outdated and no longer reflective of the current state of affairs when real-time updates are not received. This relates to privacy and data use because users could choose to remove content on Twitter, but have it remain active on Ushahidi. This counters user privacy expectations around data deletion [51].

Observation 2: Limited Anonymization Coverage Ushahidi aggregates data from various sources, including user-submitted reports (surveys), Twitter, email, and SMS [87]. The platform provides an optional anonymization control that allows platform administrators to selectively obfuscate an author's information, location and timestamps [83]. From

the documentation, it is unclear whether data anonymization features are available for information collected from sources other than Ushahidi Surveys [84]. The possible lack of anonymization features for other sources could pose a privacy risk to the reporter's identity.

Observation 3: Lack of Privacy Safeguards for Raw Data
The Ushahidi platform offers anonymization features for publishing posts to end-users. Platform admins can optionally choose to obfuscate an display fields such as author's information, location and timestamps [84]. However, the data is stored in plain text in the database without employing any data-oriented strategies (e.g., anonymization, obfuscation) to protect privacy. Storing plain-text data in the database could pose a risk [69] to reporters' privacy in certain contexts where trust is assumed: (1) malicious administrators with access to internal databases, or (2) raw data shared for secondary purposes such as research, policy-making, or compliance with law enforcement requests.

Observation 4: Use of Direct Identifiers for Unstructured Data Sources The collected data contain direct identifiers, such as the author's information. Structured data from in-platform surveys are obfuscated, while data from unstructured sources (such as email, SMS and Twitter reports) are stored and/or published without applying anonymization techniques. The use of direct identifiers in a crowdsourcing platform could single out and identify the reporter who submitted the information. Depending on the context, this may pose a serious risk or threat to the reporter (e.g., activist campaigns). Additionally, reporters' unique identifiers can be used to correlate with social networks to discover personal associations, posing a serious risk or threat not only to the reporter but also their close connections (e.g., friends or family).

Responsible Disclosure We have shared our observations with Ushahidi's security team on 12/11/23, but we have not received a response at the time of publication.

5.3 Case Study 3: DIVOC

The Digital Infrastructure for Verifiable Open Credentialing (DIVOC) is an open-source platform for countries to conduct large-scale digitized health campaigns [24]. Adopters can flexibly choose the components they want to implement and customize them to suit their needs. For example, countries can use DIVOC to establish a digital infrastructure for issuing and verifying their citizens' vaccination certificates.

DIVOC is developed and maintained by the eGov Foundation of India. It has been used by countries like India, Indonesia, Jamaica, the Philippines, and Sri Lanka to issue and verify over 2 billion COVID-19 vaccination certificates [24].

Summary of Findings DIVOC's response to Section 9(a) of the DPG questionnaire states that they do not collect PII. However, their infrastructure allows implementers to collect PII while orchestrating health campaigns. We observe that the

DPG standard is not nuanced enough to differentiate between the collection PII by the DPG or its implementers. For example, the other two DPGs we evaluated (MOSIP and Ushahidi) are used by implementers who collect PII, but the DPGs still declare they collect PII in their responses. Although DIVOC mentions implementers are responsible for protecting user privacy in their response, they also provide privacy and security best practices in their adopter-focused documentation.

5.3.1 Analysis

The DIVOC platform follows a microservice architecture and can integrate with third-party services [22]. The DFD for DIVOC is illustrated in Figure 8 in the Appendix. This illustration informed the threat elicitation process using the LIND-DUN framework [92]. At a high level, DIVOC consists of two core modules [23]. The first module is responsible for issuing, verifying, and distributing credentials (e.g., vaccination certificates). The second module monitors the performance of the health campaign by computing real-time analytics.

Countries can include several additional modules [23] in their DIVOC instance, such as a program set-up module that creates and maintains registries for credentials and facilities where these credentials are issued. A citizen portal is also available for citizens to self-register, schedule appointments with a facility, and download and verify their credentials.

Observation 1: Delegation of Responsibilities DIVOC states that they do not collect, store, or distribute PII in their response to the DPG questionnaire. However, their platform is "meant for last-mile vaccination administration and credentialing", and its implementers can collect and store PII such as name, date of birth, and identifiers like a national identity number [28]. DIVOC mentions their platform architecture prioritizes data minimalism, with "well-designed privacy & security" measures in their response. They further note that the effectiveness of the supported privacy measures depends on the individual privacy policies used by their adopters. Although they delegate the responsibility of protecting user privacy to their adopters, DIVOC provides them with privacy and security best practices to follow [25], described below.

Observation 2: Privacy Guidelines for Adopters DIVOC provides comprehensive data protection guidelines for its adopters in its documentation [25]. For example, to ensure secure data backups, DIVOC recommends implementing the principle of least privilege by restricting access to user and system information based on task requirements, as well as purging intermediate data backups and keeping full backups on separate servers after encrypting the data. It also gives recommendations on authentication and password management, access control, and platform updates. Finally, DIVOC also includes templates of user-facing privacy policies that adopters can use while running their health campaigns [26, 27].

Responsible Disclosure We had no privacy concerns to share.

6 Discussion

Our findings highlight three important points:

1. The DPG standard is not currently ensuring that DPGs offer a strong level of privacy protection. Although the intent of the privacy question on the DPG questionnaire is clearly aligned with best privacy practices, the reality is that many approved DPGs have responded to it incompletely or incorrectly, and made it through the approval process. For example, our qualitative analysis in Section 4 of DPG responses indicates that over 65% of DPGs we studied either had incomplete or vague privacy documentation; if these responses are representative of their true privacy posture, those DPGs may be vulnerable to privacy threats. Hence, the DPG questionnaire is not currently filtering out responses with a weak or incomplete description of privacy protections.

2. The current DPG standard does not collect nuanced enough information to distinguish DPGs with very different privacy profiles. Among certified DPGs, there is a broad range of levels of privacy maturity. For instance, we noted that MOSIP had implemented and documented many privacy features, whereas DIVOC chose to implement relatively fewer privacy features, leaving a significant amount of implementation to the DPG adopter. There could exist a version of the DPG standard that differentiates between these two very different models of implementation. We give one proposal for how to design such a model in Section 6.2. However, we note that the DPGA may not wish to be responsible for differentiating between DPGs of differing privacy postures, as this would require a much more in-depth analysis.

3. Should the DPGA be evaluating DPGs' privacy posture? A broader question is whether the DPGA should be tasked with evaluating or ensuring the privacy of DPGs. Currently, the DPGA may be constrained in part by the UN's report, which emphasizes the importance of privacy, and in part by the lack of clear guidance internationally on how to evaluate the privacy of software (let alone other classes of DPGs like machine learning models, datasets, etc.). Hence, it may be worth revisiting whether the DPGA's role in privacy evaluation. We discuss an alternative model in Section 6.2.

6.1 Limitations

Our methodology has some limitations, which we highlight here. First, our DPG sample is biased, including only DPGs that were certified. It would be useful to also analyze the responses of DPG candidates that were not approved, but this data is not publicly available.

Another important limitation of our methodology is that it rewards DPGs with more developed privacy documentation, regardless of how developed their privacy features are. For example, as we saw in our case studies in Section 5, DIVOC fared well in part because it did not specify many implementation details for privacy functionalities. Instead, it delegated

responsibility to solution adopters, and documented recommendations clearly in its documentation. This prevented our threat elicitation process from identifying threats in its data flows. On the other hand, MOSIP implemented (and documented) more privacy features, so it was easier to observe concrete gaps. A more complete prototype like MOSIP may require less effort from adopters, who will most likely use out-of-the-box privacy features. However, it is difficult to directly compare DPGs with differing levels of implementation.

6.2 Recommendations

6.2.1 DPG Community

(1) Do not refine the DPG questionnaire with more specific questions. A natural reaction to our findings is to attempt to revise the DPG standard to be more precise and granular about what privacy properties a DPG should satisfy. We suggest *not* pursuing such a direction. Since the DPG standard is meant to be adopted globally, building consensus around a privacy standard is likely to be politically challenging. Privacy norms are highly culture-specific [93], and we note that to the best of our knowledge, there are no true privacy standards in place that address an entire product, even among (inter-)national standards bodies; instead, the focus has been on building general-purpose *frameworks* that are very high-level, but also broadly applicable [71]. Second, privacy best practices are often technology-specific, and it is unclear how to craft a standard that encompasses the broad range DPGs (e.g., a national ID system vs. a machine learning model).

(2) Adopt a new architecture for collecting privacy evaluations. Instead of updating the DPG standard to be more comprehensive, we suggest a model that makes use of the existing ecosystem for privacy evaluation, which are themselves the products of many years of refinement and stakeholder engagement [64, 65, 92]. Our proposed model would have two tiers of privacy certification (see Figure 5).

Tier 1: Certified Privacy Impact Assessment. At the stronger tier, the DPGA would ask candidate DPGs to submit documentation attesting to the fact that they underwent a Privacy Impact Assessment (PIA) or a comparable regional variant, such as the Singaporean Data Protection Trustmark [8] from a certified provider. A PIA is an analysis of how personally identifiable information is collected, used, shared, and maintained; it involves answering a list of questions regarding data collection, retention, use, and more [64]. It provides a more fine-grained view of a product's privacy posture than the current DPG questionnaire, as studied in this work. PIAs have international adoption and are currently mandated for U.S. federal agencies by the e-Commerce Act of 2002 [36] and by the E.U. through GDPR Article 35 for all high-risk data processing activities [80]. Under our suggested process, the DPGA would collect and publish evidence of a PIA (or a comparable alternative) from an approved provider; the

DPGA can maintain a list of acceptable assessment tools and assessors e.g., [66, 67]. In addition, DPGs would upload the outcomes from the audit (answers to all questions), which should be made available on the DPGA website.

Tier 2: Self-Assessment. In the second tier, DPG candidates would submit a self-attestation that they underwent a PIA. The documentation from that process would be uploaded along with the candidate's self-attestation, so potential users can view the DPG's self-evaluated privacy posture.
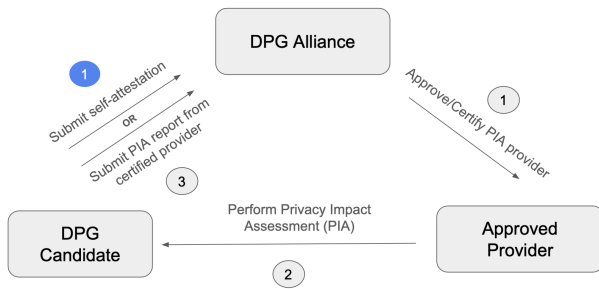
Figure 5: 3-stakeholder model to facilitate DPG privacy evaluation. The third-party assessment would involve the gray sequence of steps, whereas a self-assessment would require only the single blue step.

Under our proposed model, the DPGA would approve DPGs as long as they have accomplished one of the two. In particular, the DPGA would *not* directly evaluate, or provide their seal of approval, to DPGs' privacy postures. Evaluation of privacy documentation would be handled by the adopting entity. Note that many DPGs are not hosted services, but require an integrator (often a government) to host and run the DPG. In these cases, it is reasonable to expect a government to expend resources to evaluate the privacy posture of a piece of software before using it on constituents' data.

A potential drawback of this suggested architecture is that it increases the barrier to entry for new DPGs. However, privacy was presented in the UN's mandate as a first-order requirement of DPGs. If this is the case, it may be necessary to raise the barrier to DPG certification to ensure that DPGs are handling user data properly. We provide a stakeholder cost analysis comparing the two tiers in Appendix C.

6.2.2 Research Community

(1) Further research is needed on communicating the privacy posture of DPGs to downstream adopters. There is an active body of research on communicating the privacy posture of applications to end users [39, 43, 94]. One well-known example is privacy nutrition labels [43, 49]. These technologies may be nontrivial to apply to DPGs. Adopters of a DPG could be governments or hobbyists, and they may use the same tool for very different purposes. Hence, their privacy needs may vary significantly, so the structure that makes privacy

nutrition labels easy for users to understand may not extend easily to DPGs. Second, DPGs often limit what aspects of the system they implement, and which parts they leave to the downstream adopter. This can impact privacy in nuanced ways (as shown in our case studies in §5), and those impacts should be communicated clearly. In sum, understanding how to clearly communicate the privacy (and security) posture of a DPG is an interesting and complex question for the usable security and privacy research community.

(2) Continue to develop automated tools for dynamically evaluating the privacy posture of software. A drawback of the suggested architecture is staleness; a privacy audit typically has a short shelf life because every new feature can introduce new privacy vulnerabilities. Hence, inspired by OS-CAL [61], continued research is needed on automatically processing a codebase and extracting potential privacy vulnerabilities. While this is already a rich area of research [2, 46, 72], there is still room to make these tools usable and connect them to standardized privacy certifications.

6.3 Ethical Considerations

This study is not human subjects research, and it used only public data about products (not people). The study was not subject to review by our Internal Review Board. We followed industry-standard best practices for disclosing potential vulnerabilities to MOSIP and Ushahidi after our case studies, and gave both 60 days' notice prior to publicizing results [15].

7 Conclusion

This work provides the first large-scale study of DPGs and their privacy properties. Our results suggest that the DPG standard may benefit from revising its methodology for evaluating DPG candidates' privacy maturity. We have communicated our findings and recommendations with the DPGA, which is currently revising the DPG standard (although we are not sure in which direction). In addition to encouraging the DPGA to improve the DPG privacy certification process, we hope this study will inform future privacy-conscious initiatives, such as the emerging push for digital public infrastructure [70]. We also hope this work will encourage the usable privacy research community to explore ways of communicating privacy to third-party adopters of user-facing technologies.

Acknowledgments

References

[1] Tech monopolies and the insufficient necessity of interoperability. https://onezero.medium.com/tech-monopolies-and-the-insufficient-necessity-of-interoperability-aafba94f1eb3, 2024. (Accessed on February 14, 2024).

[2] Rafael Accorsi. Automated privacy audits to complement the notion of control for identity management. In *Policies and Research in Identity Management: First IFIP WG11. 6 Working Conference on Policies and Research in Identity Management (IDMAN'07), RSM Erasmus University, Rotterdam, The Netherlands, October 11-12, 2007*, pages 39–48. Springer, 2008.

[3] Accountability Act. Health insurance portability and accountability act of 1996. *Public law*, 104:191, 1996.

[4] Idris Adjerid, Alessandro Acquisti, Rahul Telang, Rema Padman, and Julia Adler-Milstein. The impact of privacy regulation and technology incentives: The case of health information exchanges. *Management Science*, 62(4):1042–1063, 2016.

[5] Digital Public Goods Alliance. Digital public goods registry. https://digitalpublicgoods.net/registry/.

[6] Digital Public Goods Alliance. Digital public goods standard questionnaire. https://github.com/DPGAlliance/DPG-Standard/blob/main/standard-questions.md.

[7] Paul Ashley, Calvin Powers, and Matthias Schunter. From privacy promises to privacy management: a new approach for enforcing privacy throughout an enterprise. In *Proceedings of the 2002 workshop on New security paradigms*, pages 43–50, 2002.

[8] Infocomm Media Development Authority. Data protection trustmark (dptm) certification. https://www.imda.gov.sg/how-we-can-help/data-protection-trustmark-certification. (Accessed on May 23, 2024).

[9] Peter Bendor-Samuel. Trends in rising prices affecting companies using third-party services. https://www.forbes.com/sites/peterbendorsamuel/2021/08/02/trends-in-rising-prices-affecting-companies-using-third-party-services/?sh=2977dc4827b2, August 2 2021. (Accessed on February 14, 2024).

[10] Ann Cavoukian. Privacy by design. 2009.

[11] Yan Chen, Rosta Farzan, Robert Kraut, Iman Yeckehzaare, and Ark Fangzhou Zhang. Motivating experts to contribute to digital public goods: A personalized field experiment on wikipedia. *Management Science*, 2023.

[12] Aloni Cohen. Attacks on deidentification's defenses. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 1469–1486, 2022.

[13] Asia-Pacific Economic Cooperation. Apec privacy framework. *Asia Pacific Economic Cooperation Secretariat*, 81, 2005.

[14] Forbes Tech Council. Gdpr and the 'security by compliance' mistake. https://www.forbes.com/sites/forbestechcouncil/2018/07/02/gdpr-and-the-security-by-compliance-mistake/?sh=4d0d78fecc4d, July 2 2018. (Accessed on February 14, 2024).

[15] Cybersecurity & Infrastructure Security Agency. Bod 20-01: Develop and publish a vulnerability disclosure policy | cisa. https://www.cisa.gov/news-events/directives/bod-20-01-develop-and-publish-vulnerability-disclosure-policy#:~:text=Many%20in%20the%20security%20research,the%20issue%20without%20unnecessary%20delay. (Accessed on February 09, 2024).

[16] Digital Public Goods Alliance. Digital public goods. https://digitalpublicgoods.net/digital-public-goods/. (Accessed on February 14, 2024).

[17] Digital Public Goods Alliance. Digital public goods standard. https://digitalpublicgoods.net/standard/. (Accessed on February 14, 2024).

[18] Digital Public Goods Alliance. Digital public goods repository. https://github.com/DPGAlliance/publicgoods-candidates/tree/main/digitalpublicgoods, 2024. (Accessed on February 14, 2024).

[19] Digital Public Goods Alliance. Who we are. https://digitalpublicgoods.net/who-we-are/, n.d. (Accessed on February 14, 2024).

[20] Cynthia Dwork, Adam Smith, Thomas Steinke, and Jonathan Ullman. Exposed! a survey of attacks on private data. *Annual Review of Statistics and Its Application*, 4:61–84, 2017.

[21] David Eaves, Leonie Bolte, Omayra Chuquihuara Gozalo, and Surabhi Hodigere Raghavendra. Best practices for the governance of digital public goods. *Ash Center Policy Briefs Series*, 2022.

[22] eGov Foundation. Divoc architecture. https://divoc.digit.org/platform/divoc-architecture.

[23] eGov Foundation. Divoc demo for modules. `https://divoc.digit.org/divoc-demo`.

[24] eGov Foundation. Divoc egov foundation. `https://divoc.egov.org.in/`.

[25] eGov Foundation. Divoc platform policy guidelines. `https://divoc.digit.org/community/about-project-team/platform-policy-guidelines`.

[26] eGov Foundation. Divoc privacy policy: Detailed. `https://divoc.digit.org/community/about-project-team/privacy-policy-detailed`.

[27] eGov Foundation. Divoc privacy policy: Short version for display. `https://divoc.digit.org/community/about-project-team/privacy-policy-short-version-for-display`.

[28] eGov Foundation. Divoc: What information goes into a qr code? `https://divoc.digit.org/platform/divocs-verifiable-certificate-features-2.0/what-information-goes-into-a-qr-code`.

[29] Marcello Estevao. 4 ways low-income economies can boost tax revenue without hurting growth. `https://blogs.worldbank.org/voices/4-ways-low-income-economies-can-boost-tax-revenue-without-hurting-growth`, 2024. (Accessed on February 14, 2024).

[30] Alessandro Fedele and Cristian Roner. Dangerous games: A literature review on cybersecurity investments. *Journal of Economic Surveys*, 36(1):157–187, 2022.

[31] Batya Friedman, Edward Felten, and Lynette I Millett. Informed consent online: A conceptual model and design principles. *University of Washington Computer Science & Engineering Technical Report 00–12–2*, 8, 2000.

[32] GeeksforGeeks. Levels in data flow diagrams (dfd). `https://www.geeksforgeeks.org/levels-in-data-flow-diagrams-dfd/`, 2024. (Accessed on February 14, 2024).

[33] United Nations Secretary General. Roadmap for digital cooperation. `https://www.un.org/en/content/digital-cooperation-roadmap/assets/pdf/Roadmap_for_Digital_Cooperation_EN.pdf`, 2020. (Accessed on February 14, 2024).

[34] Alison Gillwald and Anri van der Spuy. The governance of global digital public goods: Not just a crisis for africa. *GigaNet, Berlin*, 2019.

[35] Aam Digital GmbH. Aam digital. `https://aam-digital.com/`.

[36] U.S. Government. Public Law 107–347, 107th Congress. `https://www.govinfo.gov/content/pkg/PLAW-107publ347/pdf/PLAW-107publ347.pdf`, December 2022. (Accessed on February 10, 2024).

[37] Phil Gramm. Gramm–leach–bliley act. In *Vol. Public Law 106–102). Washington, DC: United States Congress*, 1999.

[38] Kendra Gray. The privacy rule: Are we being deceived. *DePaul J. Health Care L.*, 11:89, 2007.

[39] Margaret Hagen. {User-Centered} privacy communication design. In *SOUPS*, 2016.

[40] J-H Hoepman. Privacy design strategies (the little blue book). 2018.

[41] Michael Howard and David LeBlanc. Uncover security design flaws using the stride approach. *MSDN Magazine*, November 2006. (Accessed on February 14, 2024).

[42] Intersoft Consulting. Article 5 - principles relating to processing of personal data. `https://gdpr-info.eu/art-5-gdpr/`, 2023. (Accessed on November 10, 2023).

[43] Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W Reeder. A" nutrition label" for privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, pages 1–12, 2009.

[44] Brian B Kelly. Investing in a centralized cybersecurity infrastructure: Why hacktivism can and should influence cybersecurity reform. *BUL Rev.*, 92:1663, 2012.

[45] Michael Kummer, Olga Slivko, and Xiaoquan Zhang. Unemployment and digital public goods contribution. *Information Systems Research*, 31(3):801–819, 2020.

[46] Immanuel Kunz, Konrad Weiss, Angelika Schneider, and Christian Banse. Privacy property graph: Towards automated privacy threat modeling via static graph-based analysis. *Proceedings on Privacy Enhancing Technologies*, 2023.

[47] WYNNE LAM and Bruce Lyons. Data protection legislation and investment incentives when consumers are loss averse. 2019.

[48] Marc Lelarge and Jean Bolot. Economic incentives to increase security in the internet: The case for insurance. In *IEEE INFOCOM 2009*, pages 1494–1502. IEEE, 2009.

[49] Tianshi Li, Kayla Reiman, Yuvraj Agarwal, Lorrie Faith Cranor, and Jason I Hong. Understanding challenges for developers to create accurate privacy nutrition labels. In *CHI*, 2022.

[50] Emma L. Slade Marijn Janssen, Nripendra P. Rana and Yogesh K. Dwivedi. Trustworthiness of digital government services: deriving a comprehensive theory through interpretive structural modelling. *Public Management Review*, 20(5):647–671, 2018.

[51] Mohsen Minaei, Mainack Mondal, and Aniket Kate. Empirical understanding of deletion privacy: Experiences, expectations, and measures. In *USENIX Security*, pages 3415–3432, 2022.

[52] MOSIP. A digital public good for identity. `https://mosip.io/#1`. (Accessed on February 14, 2024).

[53] Mosip Documentation. Mosip Documentation - Administration. `https://docs.mosip.io/1.2.0/module s/administration#what-is-deactivation-of-a -resource`. (Accessed on February 14, 2024).

[54] Mosip Documentation. Mosip Documentation - Anonymous Profiling Support. `https://docs.mosip.io/1. 2.0/id-lifecycle-management/anonymous-pro filing-support`. (Accessed on February 14, 2024).

[55] Mosip Documentation. Mosip Documentation - Collab Pre-registration Guide. `https://docs.mosip.io/1. 2.0/collab-getting-started-guide/collab-p re-registration-guide`. (Accessed on February 14, 2024).

[56] Mosip Documentation. Mosip Documentation - ID Authentication. `https://docs.mosip.io/1.2.0/id-a uthentication`. (Accessed on September 22, 2023).

[57] Mosip Documentation. Mosip documentation - overview. `https://docs.mosip.io/1.2.0/over view`. (Accessed on February 14, 2024).

[58] Mosip Documentation. Mosip documentation - registration client home page. `https://docs.mosip.io/1.2. 0/modules/registration-client/registration -client-home-page#new-registration`. (Accessed on September 22, 2023).

[59] Anit Mukherjee and Shankar Maruwada. Fast-tracking development: A building blocks approach for digital public goods. *Center for Global Development. https://www. cgdev. org/sites/default/files/fast-tracking-development-digital-publicgoods. pdf*, 2021.

[60] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125. IEEE, 2008.

[61] National Institute of Standards and Technology. Oscal - open security controls assessment language. `https:// pages.nist.gov/OSCAL/`, n.d. (Accessed on February 14, 2024).

[62] Brian Nicholson, Petter Nielsen, Johan Ivar Sæbø, and Ana Paula Tavares. Digital public goods for development: A conspectus and research agenda. In *International Conference on Social Implications of Computers in Developing Countries*, pages 455–470. Springer, 2022.

[63] Brian Nicholson, Petter Nielsen, Sundeep Sahay, and Johan Ivar Sæbø. Digital public goods platforms for development: The challenge of scaling. *The Information Society*, 38(5):364–376, 2022.

[64] Department of Homeland Security. Privacy office official guidance for privacy impact assessments. `https: //www.dhs.gov/sites/default/files/publicat ions/privacy_pia_guidance_june2010_0.pdf`. (Accessed on February 10, 2024).

[65] National Institute of Standards and Technology. Nist privacy framework. `https://www.nist.gov/priva cy-framework/privacy-framework`, Jan 2024.

[66] OneTrust. Pia and dpia automation. `https://www.on etrust.com/products/pia-and-dpia-automatio n/`. (Accessed on May 23, 2024).

[67] Osano. Data privacy assessment tool. `https://ww w.osano.com/products/privacy-assessments`. (Accessed on May 23, 2024).

[68] Cliodhna O'Connor and Helene Joffe. Intercoder reliability in qualitative research: debates and practical guidelines. *International journal of qualitative methods*, 19:1609406919899220, 2020.

[69] Jordan Pearson. The breach of a face recognition firm reveals a hidden danger of biometrics, May 2024.

[70] United Nations Development Programme. Digital public infrastructure. `https://www.undp.org/digital/d igital-public-infrastructure`. (Accessed on May 18, 2024).

[71] Privacy Ref. Choosing a privacy framework. `https: //privacyref.com/blog/choosing-a-privacy-f ramework/`, 2024. (Accessed on February 14, 2024).

[72] Jenni Reuben, Leonardo A Martucci, and Simone Fischer-Hübner. Automated log audits for privacy compliance validation: a literature survey. *Privacy and Identity Management. Time for a Revolution? 10th IFIP WG 9.2, 9.5, 9.6/11.7, 11.4, 11.6/SIG 9.2. 2 International Summer School, Edinburgh, UK, August 16-21, 2015, Revised Selected Papers 10*, pages 312–326, 2016.

[73] John P Rosson, Mason J Rice, Juan Lopez Jr, and Robert David Fass. Incentivizing cyber security investment in the power sector using an extended cyber

insurance framework. *Homeland Security Affairs*, 15, 2019.

[74] Johan Ivar Sæbø, Brian Nicholson, Petter Nielsen, and Sundeep Sahay. Digital global public goods. *arXiv preprint arXiv:2108.09718*, 2021.

[75] Swarag Sharma and Jevitha KP. Security analysis of oauth 2.0 implementation. In *2023 Innovations in Power and Advanced Computing Technologies (i-PACT)*, 2023.

[76] Sean Sirur, Jason RC Nurse, and Helena Webb. Are we there yet? understanding the challenges faced in complying with the general data protection regulation (gdpr). In *Proceedings of the 2nd International Workshop on Multimedia Privacy and Security*, pages 88–95, 2018.

[77] Matthias Stürmer, Markus Andreas Tiede, Jasmin Myriam Nussbaumer, and Flurina Wäspi. On digital sustainability and digital public goods. 2023.

[78] Latanya Sweeney. Simple demographics often identify people uniquely. *Health (San Francisco)*, 671(2000):1–34, 2000.

[79] Ana Paula Tavares, Edgar Whitley, Liv Marte Nordhaug, Johan Sæbø, Malavika Raghavan, PK Senyo, and Silvia Masiero. Digital public goods and vulnerable populations. 2023.

[80] European Union. Article 35 of gdpr - data protection impact assessment. `https://gdpr.eu/article-35-impact-assessment/`. (Accessed on February 10, 2024).

[81] United States Government. Children's online privacy protection rule ("coppa"). `https://uscode.house.gov/view.xhtml?req=granuleid%3AUSC-prelim-title15-section6501&edition=prelim`. (Accessed on February 13, 2024).

[82] Ushahidi. Ushahidi. `https://www.ushahidi.com`.

[83] Ushahidi. Data obfuscation. `https://www-admin.ushahidi.com/support/data-obfuscation`, n.d. (Accessed on February 14, 2024).

[84] Ushahidi. Post types. `https://www-admin.ushahidi.com/support/post-types#what-exactly-is-a-survey`, n.d. (Accessed on February 14, 2024).

[85] Ushahidi. Ushahidi deployments. `https://www.ushahidi.com/in-action/deployments/`, n.d. (Accessed on February 14, 2024).

[86] Ushahidi. Ushahidi platform developer documentation: Architecture. `https://docs.ushahidi.com/platform-developer-documentation/tech-stack/architecture`, n.d. (Accessed on February 14, 2024).

[87] Ushahidi. Ushahidi platform user manual: Configuring data sources. `https://docs.ushahidi.com/platform-user-manual/3.-configuring-your-deployment/3.4-data-sources`, n.d. (Accessed on February 14, 2024).

[88] Ushahidi. Ushahidi platform user manual: Configuring twitter data source. `https://docs.ushahidi.com/platform-user-manual/3.-configuring-your-deployment/3.4-data-sources/3.4.7-twitter`, n.d. (Accessed on February 14, 2024).

[89] Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

[90] Richmond Y Wong, Andrew Chong, and R Cooper Aspegren. Privacy legislation as business risks: How gdpr and ccpa are represented in technology companies' investment risk disclosures. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–26, 2023.

[91] David Wright and Paul De Hert. *Enforcing privacy: regulatory, legal and technological approaches*, volume 25. Springer, 2016.

[92] Kim Wuyts and Wouter Joosen. Linddun privacy threat modeling: a tutorial. *CW Reports*, 2015.

[93] Abbas Zabihzadeh, Mohammad Ali Mazaheri, Javad Hatami, Mohammad Reza Nikfarjam, Leili Panaghi, and Telli Davoodi. Cultural differences in conceptual representation of "privacy": A comparison between iran and the united states. *The Journal of social psychology*, 159(4):357–370, 2019.

[94] Sebastian Zimmeck, Eliza Kuller, Chunyue Ma, Bella Tassone, and Joe Champeau. Generalizable active privacy choice: Designing a graphical user interface for global privacy control. *PETS*, 2024.

A Codebooks

Table 1: Codebook for Proposed Protection Mechanisms.

Code	Definition
Access Control	The nominee proposes the enforcement of access restrictions to the solution and/or personal data based on predefined rules and policies.
Commercial/In-house Tools	The nominee proposes the use of commercial or in-house tools for the protection of personal data.
Data Strategies	The nominee proposes data-oriented protection strategies (e.g., minimization, anonymization) as a protective measure for personal data.
Notify 3rd Party Data Sharing	The nominee describes the sharing of personal data with third parties for secondary use.
Security Upgrades/Patches	The nominee proposes taking responsibility for performing regular security updates and patches to protect personal data.
Vulnerability Testing	The nominee proposes taking responsibility for performing regular vulnerability scans to ensure protection of personal data.
Governance Processes/Audits	The nominee proposes taking responsibility for establishing and/or adhering to a governance process to ensure the protection of personal data.
User Control	The nominee proposes implementing various privacy controls to empower users in expressing their privacy preferences effectively (e.g., user consent, data deletion requests)
Data Storage	The nominee proposes secure data storage solution(s) to ensure protection of personal data.
Strong Passwords	The nominee proposes safeguarding access to personal data by implementing robust password requirements.

Table 2: Codebook for Provided Supporting Material

Code	Definition
No Documentation Submitted	The nominee has not submitted any documentation or references, or has submitted expired links, for review.
Security/Privacy Docs	The nominee has shared a link to the solution's security and privacy documentation, which comprises either detailed or high-level information about implementation.
Compliance Docs	The nominee has shared a link to the solution's compliance documentation (e.g., GDPR).
Cookie Policy	The nominee has shared a link to the solution's cookie practices.
Privacy Policy	The nominee has shared a link to the solution's privacy practices. Note: while this is a step towards the right direction, it is still not sufficient.

Table 3: Codebook for Overall Response Quality

Code	Definition
Security-related Privacy	Discusses security-related privacy controls, such as encryption and access control, without discussing any data-oriented strategies.
Security Only	Focuses only on security measures, without addressing any privacy-related strategies, despite PII being collected, stored, and/or processed
Partially Addresses Privacy	Addresses certain aspects of privacy but may not cover all aspects comprehensively.
Unclear PII Collection	Nominees lack a clear understanding of what personally identifiable information (PII) entails. Responses are either incorrect or unclear.
Clarifies Data Ownership	Emphasis that solution developers (nominee) do not claim ownership of any data collected and/or processed by the solution. Places the burden of privacy on solution implementers, neglecting the fact that privacy-by-design principles should have been incorporated during development.
Lack of Specificity	Mentions vague terms without explaining the solution's function/capabilities or proposes privacy-protecting solutions without providing specific implementation details.
Downplaying Risks	Clarifies that the nominee do not collect data themselves and explicitly state that they do not own any of the data. This clarification may be intended to downplay potential risks associated with protecting user data.
Compliance Implies Protection	Claims compliance with data protection regulations such as GDPR; it does not necessarily indicate that privacy-protecting strategies have been implemented.
Inconsistent Answer	Response provided directly contradicts the answers given to other questions.
Does Not Answer Question	Incorrect response that does not answer the question.

Table 4: Codebook for Privacy Component Analysis

Code	Definition
Regulatory Efforts	The nominee describes privacy related compliance efforts such as self regulation, enforcement mechanisms, privacy documentation and awareness campaigns.
Notice and Consent	Nominees describes user notice and consent mechanisms.
Data Collection Limitation	Nominee ensures that the system only collects data required for the intended purpose and as for long as necessary.
Data Use Limitation	Nominee ensures that the system only processes data needed to satisfy the intended purpose and describe strategies to do so.
User Choice	Nominee ensures that users provided with appropriate and user-friendly choices in relation to collection, use, transfer and disclosure of their personal information.
Data Accuracy	Nominee ensures that data is accurate and up-to date.
Security Safeguards	Nominee describes security safeguards to protect user data.
Privacy by Design	Nominee addresses Privacy by design (PbD) strategies such as anonymization and privacy preserving defaults.

B   Data Flow Diagrams for Case Studies
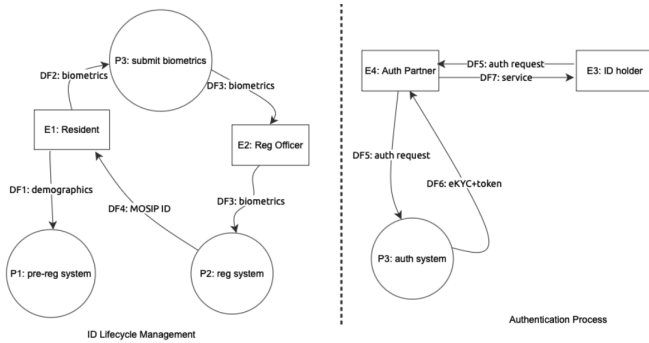
Figure 6: MOSIP's Data Flow Diagram
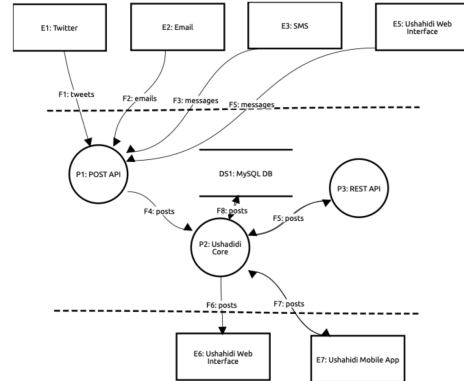
Figure 7: Ushahidi's Data Flow Diagram
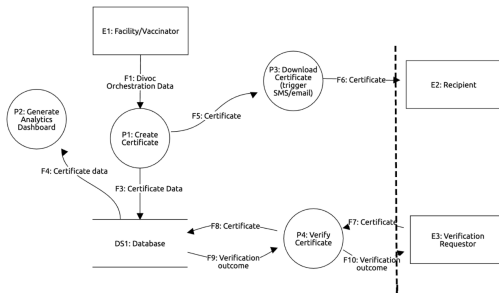
Figure 8: DIVOC's Data Flow Diagram

C   Cost Analysis

Choosing between the two tiers in our proposed model will depend on the resources available to the DPGA and the DPG candidate. Self-attestation could impose a burden on DPG candidates, many of which lack privacy and/or compliance teams. Obtaining a certification from approved providers requires less time but possibly more money for DPG candidates, depending on the time cost of completing the assessment and financial compensation of DPG contributors. While the proposed model would increase the barrier to DPG certification, we believe basic privacy assessments should be a minimum requirement for organizations handling PII. We summarize the stakeholder cost analysis for our proposed model in Table 5.

Table 5: Stakeholder cost analysis of the online cost (i.e., during DPG certification) of the two privacy certification tiers of our proposed model. Arrows indicate the change in resources compared to what is currently needed.

Proposed Strategy	Stakeholder	Time	Money	Overall Effort
Option 1: PIA by Certified Provider	DPGA	↓	↓	↓
	Approved Provider	↑	↓	↑
	DPG Cand.	↓	↑	↑
Option 2: Self Attestation	DPGA	↓	↓	↓
	DPG Cand.	↑	↑	↑

Well-intended but half-hearted: Hosts' consideration of guests' privacy using smart devices on rental properties

Sunyup Park
Univ. of Maryland,
College Park

Weijia He
Dartmouth College

Elmira Deldari
Univ. of Maryland,
Baltimore County

Pardis Emami-Naeini
Duke University

Danny Yuxing Huang
New York University

Jessica Vitak
Univ. of Maryland,
College Park

Yaxing Yao
Virginia Tech

Michael Zimmer
Marquette University

Abstract

The increased use of smart home devices (SHDs) on short-term rental (STR) properties raises privacy concerns for guests. While previous literature identifies guests' privacy concerns and the need to negotiate guests' privacy preferences with hosts, there is a lack of research from the hosts' perspectives. This paper investigates if and how hosts consider guests' privacy when using their SHDs on their STRs, to understand hosts' willingness to accommodate guests' privacy concerns, a starting point for negotiation. We conducted online interviews with 15 STR hosts (e.g., Airbnb/Vrbo), finding that they generally use, manage, and disclose their SHDs in ways that protect guests' privacy. However, hosts' practices fell short of their intentions because of competing needs and goals (i.e., protecting their property versus protecting guests' privacy). Findings also highlight that hosts do not have proper support from the platforms on how to navigate these competing goals. Therefore, we discuss how to improve platforms' guidelines/policies to prevent and resolve conflicts with guests and measures to increase engagement from both sides to set ground for negotiation.

1 Introduction

Digital platform mediated short-term rentals (STRs), such as Airbnb and Vrbo, have become increasingly popular over the last decade. With the popularity and diversity of smart home devices (SHDs), STR hosts are increasingly using SHDs to add convenience for guests and to monitor the property's and guests' safety remotely [3,12,41,55,57]. The increased use of SHDs in STRs, however, raises privacy concerns that ranges from interpersonal entities' monitoring and surveillance, to people's data being collected, stored, and shared with institutional entities such as device manufacturers, law enforcement, and third-parties [9, 13, 15, 27, 30, 35, 36, 45, 54, 55].

Research has shown that Airbnb guests are uncomfortable with devices that could potentially monitor them [36, 54]. In fact, STRs guests may have unique privacy expectations, especially when compared to those in traditional hotels. STRs provide an unique "feeling of home" to its customers [63], and people expect greater privacy at homes than any other places [16]. Therefore, guests may have greater expectations of privacy in STRs than in hotels. Additionally, STRs are generally managed by individuals (e.g., hosts) rather than corporations, indicating a possible transfer of legal responsibility [**?**], which may further enhance guests' privacy concerns.

Meanwhile, Airbnb hosts express little or no concern about guests' privacy when using SHDs on their property [15]; rather, their concerns about privacy pertained to guests accessing hosts' data. At the same time, STR hosts are incentivized to accommodate guests' privacy expectations to ensure their positive experience. One way to do this is for guests to negotiate their privacy needs with hosts [54] and address any tension between hosts' goals of using SHDs and guests' values of privacy. What is not clear, however, is whether hosts feel willing and able to engage in such negotiation.

In this paper, we focus on investigating hosts' perspective on if and how they negotiate privacy with guests. Privacy negotiation involves multiple stakeholders trying to reach a consensus regarding data collection practices [54]. Majority of prior work focuses on guests' privacy needs [36,54]. While Dey et al. [15] explored Airbnb hosts' motivation to use SHDs, we still lack knowledge about hosts' current practices around SHD usage, especially in terms of managing [1] and disclosing SHDs. Therefore, we ask the following research questions:

RQ1: How do short-term rental (STR) hosts use smart home devices (SHDs) in their rental properties?

---

[1]In this paper, we define smart home device management as managing accounts, reviewing and deleting data, and granting control of devices.

RQ2: How do STR hosts manage SHDs on their properties?

RQ3: How do STR hosts communicate about their SHDs with guests?

To answer our research questions, we conducted an exploratory interview study with 15 STR hosts (e.g., Airbnb/Vrbo) about their usage of SHDs on their rental properties, how they manage and communicate with guests about their devices, and how, if at all, they consider guests' privacy when making decisions related to SHDs. Aligning with prior work [15, 36], we found that hosts use SHDs for safety and security purposes, which inevitably monitor their guests. Contrary to prior research [15, 36], however, we found that hosts take guests' privacy into consideration, albeit in limited ways. Hosts consider guests' privacy when deciding which devices to use and where to locate them, logging out of guests' accounts, limiting monitoring and control during guests' visits, and disclosing their devices to guests. We also found that hosts rarely, if ever, review or delete data; they provide limited control options to guests; and they do not disclose all SHDs.

Our paper contributes to the existing literature on privacy negotiation among multiple stakeholders in three key ways:

- We highlight hosts' conflicting needs in protecting their STRs versus protecting guests' privacy.
- We describe hosts' (limited) actions to ease guests' privacy concerns, especially in managing SHDs' data.
- We provide recommendations to improve platforms' policies/guidelines and design features to prevent and facilitate privacy negotiation between hosts and guests.

2   Related Work

2.1   Multi-user interactions in smart homes

A smart home is a multi-user environment that involve primary users and non-primary users (e.g., alternate primary users, secondary users, and guests) based on different roles and usage scenarios [20, 25, 28, 34, 52, 59]. Primary users are those involved in purchasing, installing, using, and managing SHDs [20], while non-primary users are those who are less involved in managing SHDs, but focus on using the SHDs managed by the primary user [28]. Bystanders are an important subset of non-primary users [60], and are users who "happen to" use SHDs (also referred to as "passenger users" by [28] and "incidental users" by [13]). Research has found that when primary and non-primary users have different ideas about privacy [7, 28, 39, 40], there could be tensions and conflicts, such as passengers' concerns about the device purpose and potential surveillance and monitoring [13, 28, 29].

In the short-term rental (STR) context, we consider hosts as primary users because they purchase, install, use, and manage SHDs for their rental properties. Likewise, we consider guests bystanders because they use or are exposed to data collection by SHDs that hosts have in their STRs. However, unlike previous studies that address traditional home setups [13, 28, 52, 59, 60], stakeholders in STRs are based on a transactional relationship, incentivizing negotiation. Guests can always pick a different listing if they are unhappy with the property (e.g., hosts' usage of SHDs), while hosts are motivated to attract more guests. Therefore, it gives guests the power to negotiate, making it an ideal setting to study how people negotiate their privacy preferences.

2.2   Smart rentals and privacy

Privacy is an issue for both hosts and guests in the STR context. For example, hosts are concerned about privacy when their identities are disclosed through their public profiles [50] while guests concerned about smart devices such as hidden cameras, [11, 21, 58], general smart cameras [13, 17, 42], and smart speakers [13]. In fact, Schutte [45] found that guests were less satisfied staying in rental properties with SHDs and identified privacy as one of the reasons.

As one of the most popular STR platforms [49], Airbnb highlights the tension and conflicts between hosts and guests. From the hosts' perspective, privacy was rarely considered and if it was, it was about hosts' own privacy (e.g., guests accessing hosts' information through SHDs) [15]. From the guests' perspective, they were concerned about being monitored by SHDs and the lack of control they have with the devices, and thus, had different views on information sharing (e.g., Airbnb hosts wanted to access guests' search history, but guests were uncomfortable with sharing that information) [36]. Even for less privacy-invasive devices (e.g., thermostats), hosts and guests had conflicts about how much control they want and related access to data [35]. Wang et al. [54] further identified specific devices (e.g., security cameras, voice assistants, motion sensors) that made guests uncomfortable and suggested privacy negotiation with hosts as a possible way to lessen guests' privacy concerns.

Our study complements these findings by providing insights into how hosts use their SHDs, including data management and disclosing their devices, and how—if at all—they negotiate with guests regarding SHDs in their properties.

2.3   Smart home device and data management

In theory, smart home users can mitigate their privacy concerns by engaging in privacy-protecting behaviors, such as adjusting the location of the devices [61], avoiding using certain functions [51], receiving notifications [31], or in some cases, avoid purchasing them in the first place [23]. Less intuitive and therefore uncommon is to take technical actions, such as changing passwords and/or using two-factor authentication [51], turning-off microphones, deleting video recordings and/or behavior logs [56]. In fact, Jin et al. [26] reported that less than 1% of their respondents take technical actions to manage their smart speakers. In practice, even

smart home power users find it difficult to engage in technical measures to protect their privacy in smart homes [33]. Other research has also investigated data access and control with institutional entities (e.g., manufacturers, advertisement companies, government) [1, 4, 5] and data-sharing behaviors among interpersonal relationships (e.g., family members, domestic workers, guests) [1, 4, 18, 22, 37, 38].

In the context of STRs, a few studies have investigated ways to mitigate guests' privacy, such as building a smart home interface based on local network instead of cloud-based [19], or using blockchain technology to lessen the privacy threats in home sharing economy [24]. Marky et al. [37] further suggested that guests value the feedback of privacy protection status from the hosts and privacy protection should foster collaboration between guests and hosts. The STR context, especially from the hosts' perspective, is uniquely different from other contexts (e.g., visiting friends) and may introduce new interactions and reactions to SHDs. Our work extends the prior work with an emphasis on the hosts' perspective.

3 Method

To answer our research questions (RQs), we conducted online interviews with 15 short-term rental (STR) hosts (e.g., Airbnb/Vrbo). Our study was approved by the first author's institution's IRB.

3.1 Recruitment

We used a short screening survey to recruit participants based on two criteria: (1) they are currently hosting one or more STR properties, and (2) they are currently using or interested in using smart home devices (SHDs) in their STR properties. We initially targeted Airbnb hosts but faced significant difficulties recruiting them. At the same time, we found out that many people cross-host on Airbnb and Vrbo. Therefore, we decided to expand and recruit Vrbo hosts as well.

Recruiting STR hosts was extremely difficult due to the exclusiveness of the community. We recruited through word-of-mouth, social media, and online groups targeted to Airbnb hosts (e.g., Airhost forum, subreddits for Airbnb hosts, and Facebook groups). After facing recruitment challenges, we additionally recruited through Craigslist, Airbnb host meetups, and posted flyers in Airbnb-dense areas. Finally, we also used a snowballing method to recruit additional participants by asking for referrals, either in their surveys or interviews. We had a total of 139 screener survey respondents. After filtering out bots and invalid STR accounts, we contacted 73 potential participants; 46 replied, and we were able to schedule 15 participants. This sample size is above average (12) for the CHI community [8]. In addition, we rigorously validated STR hosts by (1) asking for their STR profiles in the survey, (2) sending a private message to their STR profiles to validate their account, and/or (3) matching profile pictures and

descriptions with participants' survey response and interview, to ensure our data quality. Participants who completed the interviews were compensated with a US$50 gift card. Participants who introduced other participants were compensated with an additional referral fee (USD1 per referral).

3.2 Data collection

Data collection started in August, 2023 until January, 2024. We conducted two pilot interviews to revise and refine our study protocol. For example, we found that the pilot participant who was interested in using SHDs also provided valuable insights, therefore decided to recruit both hosts who use or are interested in using SHDs (e.g., P2, P12). Pilot interviews are not included in the data analysis.

3.2.1 Interview protocol

Each Zoom interview was recorded. The interviews lasted 52 minutes on average. Interview questions were divided into four sections. First, we provided our definition of SHDs – *household items that are connected to the Internet or a home network to enhance functionality, connectivity, and efficiency within the home*–and asked about their motivations for using SHDs. Second, we asked our participants about their experiences using SHDs on their STRs, focusing on how they manage their devices and if there were any challenges in managing their devices. Third, we asked our participants about their perceptions and practices of disclosing SHDs to guests. We then introduced participants' STR platforms' guidelines and policies regarding SHD disclosure and asked about their familiarity and perceptions. The final set of questions covered various privacy considerations with SHD use. We first asked participants about general concerns related to SHDs and potential issues stakeholders might face. If, until this point, participants did not mention privacy concerns, which was rare, we introduced an example of hosts and guests conflict when using smart speakers. We asked our participants about their thoughts on this situation. We ended our interview by asking participants to brainstorm resolving conflicts around using SHDs from multiple stakeholders' perspectives The interview protocol is provided in appendix A.

3.2.2 Participant information

As shown in Table 1, our participants consist of 13 Airbnb and 2 Vrbo hosts, whose hosting experience, numbers and types of properties, and experiences with SHDs vary. Among SHD users (n=13), the number of devices ranged from two to ten. Our participants ranged from 25 to 65+ years old, and identified as female (n=8) and male (n=7). Most participants identified as white (n=11), and most (n=11) had at least a bachelor's degree.

ID	Age	Gender	Platform	Hosting Time	Types of property	Types of SHDs used (or want to use)	Familiarity with SHDs
P1	25-34	Male	Airbnb	before 2018	primary residence	Speakers/Voice Assistants(VAs), lights, TVs, cameras, alarms	extremely familiar
P2*	65+	Male	Airbnb	since 2022	secondary residence	(Lights, thermostats)	slightly familiar
P3**	45-54	Male	Airbnb	since 2023	investment property	Lights, thermostats, TVs, doorbells, door locks, routers, appliances, switches, alarms	very familiar
P4	35-44	Male	Airbnb	since 2021	secondary residence	Speakers/VAs, thermostats, TVs, door locks, cameras, appliances, garage doors, switches, sensors, alarms	extremely familiar
P5	45-54	Male	Airbnb	since 2022	investment property	TVs, doorbells, door locks	extremely familiar
P6	35-44	Female	Airbnb	since 2021	primary residence	Thermostats, TVs, vacuums, sensors, alarms	moderately familiar
P7	35-44	Male	Airbnb	since 2022	primary residence	Speakers/VAs, switches	moderately familiar
P8	35-44	Female	Airbnb	since 2021	primary residence	Thermostats, TVs, door locks, cameras, routers	extremely familiar
P9	25-34	Male	Airbnb	since 2021	primary residence	TVs, ACs	moderately familiar
P10	65+	Female	Vrbo	before 2018	secondary residence	Speakers/VAs, thermostats, TVs, window solutions	extremely familiar
P11	35-44	Female	Airbnb	since 2020	secondary residence	Speakers/VAs, TVs, door locks, cameras	very familiar
P12*	45-54	Female	Airbnb	before 2018	secondary residence	(Speakers/VAs, lights, thermostats, TVs, doorbells, door locks, cameras, alarms)	moderately familiar
P13	55-64	Female	Vrbo	before 2018	secondary residence	Speakers/VAs, TVs	very familiar
P14	25-34	Female	Airbnb	since 2021	primary residence	Speakers/VAs, thermostats, TVs, vacuums, doorbells, door locks, cameras, switches, sensors, alarms	very familiar
P15	35-44	Female	Airbnb	since 2019	primary residence	TVs, cameras	moderately familiar

* The participant does not have SHDs on their Airbnb property currently but is interested in using them.
** P3's interview was not transcribed because P3's recording was lost. We created a detailed memo of P3's session for analysis and write-up, but we did not quote him anywhere.

Table 1: Participant Information.

3.3 Data analysis

Our recruitment took 6 months in total. To ensure the progress, we analyzed our data alongside data collection. We transcribed the audio recordings using *Rev.ai* and manually cross-checked the transcriptions with the recordings for quality assurance. We then imported the transcriptions into *Atlas.ti* for qualitative coding. P3's recordings were lost due to technical issues. However, we took detailed notes during the interview and used them to validate the themes. Thus, similar to Koshy et al. [28], we did not discard P3 from our study. The three lead authors conducted multiple iterative rounds of coding, following coding guidelines by Saldaña [44]. First, we applied structural coding (i.e., building codes based on the interview protocol) and produced 15 initial codes. Next, we selected three transcripts with rich data and applied open coding based on the initial codes to expand on the codes, producing 50 codes. Last, we distributed the transcripts so that at least two researchers coded each transcript and produced 95 final codes. During this process, we met multiple times to resolve any disagreements and reach consensus on codes. The final codebook is provided in appendix B.

Data collection continued until we determined saturation had been achieved; upon hearing no new attitudes or experiences in our final two interviews, we determined that additional data collection was unlikely to yield additional insights [32]. After coding all transcripts, we selected codes that were relevant to answering our RQs and conducted thematic analysis of each, generating analytic memos [44]. The

selected codes for each RQ are:

RQ1: STR property description, types and location of SHDs, motivations for using or not using SHDs, reasons for device purchase and usage.

RQ2: SHD management (accounts, access, manual operations), (dis)advantages of using SHDs (confusion, complications, failures).

RQ3: Codes related to STR guidelines/policies (familiarity, perceptions, needs/wants), disclosure practices (perceptions, considerations, preferences) and resolving conflicts (potential conflicts, willingness to negotiate).

Given the qualitative nature of this paper, we refrain from reporting the exact number of participants for each theme. Instead, we use the following terminologies when reporting our results: few (0-25%), some (25-45%), about half (45-55%), many/most (55-75%), and almost all (75-100%). This is similar to other qualitative studies (e.g., [7, 16, 62, 64]).

3.4 Limitations

We faced significant challenges recruiting STR hosts, partially because we aimed to verify that our participants were actual hosts. Therefore, although we have limited participant numbers and diversity, we were able to capture real hosts' experiences. Social desirability bias [6] can happen in interview studies. We tried mitigating them by avoiding using languages related to privacy during recruitment and mentioning privacy before participants mentioned it. Instead, we prompted our

participants to think about privacy by asking the benefits and drawbacks of using SHDs on their STRs or introducing a situation where conflicts can arise between the host and the guest because of privacy concerns.

4 Findings

This section is organized based on our three research questions, which cover hosts' smart home device (SHD) usage practices in their short-term rentals (STRs), their device management practices, and how they communicate with guests regarding SHDs.

4.1 Hosts' usage of smart home devices

4.1.1 Types and locations of SHDs

Figure 1 shows the types and locations of SHDs used (or wanted to use) by hosts. Smart thermostats and smart speakers ($n = 10$) were the most frequently used (or wanted to use), followed by streaming devices, smart cameras, and smart door locks ($n = 9$). These devices were located in various spaces in the property, spanning from private spaces (i.e., guests only) to shared spaces (i.e., hosts and guests) to public spaces.

Guests' private spaces can range from a private room in a shared property to an entire property and include spaces such as the bedroom, living room, and the kitchen. Private spaces had the most SHDs, with entertainment devices such as smart speakers and smart TVs being most popular. Smart cameras, on the other hand, were rarely placed in guests' private spaces, which is in line with the recent update to Airbnb's guidelines banning indoor cameras [43]. Smart thermostats and smart alarms were placed in shared spaces where both hosts and guests have access. These devices were often considered as less-privacy invasive. SHDs for safety and security purposes (e.g., smart door locks, smart cameras) were placed in public spaces (e.g., front/back doors, yard).

The majority of our participants' STRs was either their primary or secondary residence, with only two participants explicitly identifying their property as an investment. Interestingly, participants whose STRs were secondary or an investment were more willing to incorporate SHDs than those whose STRs were their primary residence. This strategic approach to adopting SHDs on their property was aimed at enhancing property safety and security through remote control, particularly for hosts who lived far from the property.

4.1.2 Motivations for using SHDs in STRs

One of the main reasons our participants use SHDs on their STRs was to monitor their properties and guests to ensure security and safety. This was especially important for participants whose rental property was not their primary residence and, therefore, needed to use SHDs to manage their properties

remotely. These participants used their smartphone apps to control the devices (e.g., door locks, lights, thermostats) on the property remotely, reducing the need to be physically on-site. For example, smart cameras, doorbells, and door locks were used to ensure the number of guests was correct, given that guests often bring more people. P4 stated, *"sometimes people will show up and bring 10 more guests than they said they were going to bring, so we have some security cameras."* Among the devices used for monitoring, our participants were most cautious about using smart cameras, especially where they placed them. P1 stated, *"So for the camera, I think that's the most invasive smart home device that we have. It was important to me to put it somewhere where it's only, where its main function as a security device is most clearly limited. So that's why it faces the front door. And is only triggered when the front door opens, or when there's activity at the front door."* Some of our participants, unfortunately, experienced theft, damage, and other violations of house rules (e.g., smoking) and decided to place smart cameras indoors. P11 placed a smart camera in her living room, stating, *"I'm not sure if it's fine or not, but it's just, it's my property, and it, there are things that are very expensive, and as much as they could pay for it, there's a lot of effort that it takes getting the stuff back again and put it in place."* Our study was conducted before Airbnb updated its guidelines to ban indoor cameras.

4.1.3 Reasons hosts *don't* use SHDs in rental properties

While SHDs were widely used, some participants also explicitly highlighted reasons why they avoided certain devices.

The primary concern revolves around the potential violation of guests' privacy, which could hinder their comfort during their stay at the rental property. Participants emphasized their reluctance to monitor guests inside the property (e.g., avoiding using smart cameras in private spaces). For example, P8 stated, *"I would never put one [smart cameras] inside, obviously, like you'd get kicked off the platform as, and it's super creepy."* Additionally, participants expressed concerns about potential privacy breaches and discomfort for guests. They worried that smart speakers might encroach on guests' privacy by listening to conversations and raised concerns about data collection by these devices. For instance, P8 stated: *"I wouldn't use one of those [smart speakers and voice assistance] in my unit probably because of audio recording. Um, and I think that's kind of what I was hinting at earlier about like tablets or speakers or whatever they are that, um, do record. I'm, I probably side with a guest on that."* In this case, hosts' threat model includes manufacturers and companies accessing guests' data. However, hosts themselves can also be a threat, for example, using "drop in" modes to listen to guests [2]. In the next section(section 4.2.3), we elaborate more on those cases where the host could threaten guests' privacy when using SHDs.

Technical difficulties and the cost of the devices are other

Figure 1: Types and locations of SHDs used by hosts. Private spaces are those areas accessible only to guests. Shared spaces refer to locations utilized by both guests and hosts. Public spaces are areas accessible to the general public.

	Smart speakers	Smart TVs	Smart thermostat	Stand-alone streaming	Smart light bulbs	Smart home appliances	Smart alarms	Smart window	Smart cameras	Smart home sensors	Smart vacuum	Smart hubs and routers	Smart door locks	Smart doorbell	Smart garage doors
Private space only for guest	8	7	5	5	4	3	2	2	1	1	1	0	0	0	0
Shared spaces with host and guest	2	0	5	4	2	2	4	1	2	2	0	1	0	0	0
Public space open to public	0	0	0	0	0	0	0	0	6	0	0	0	9	4	1

reasons for not using certain SHDs on the property, as some participants express challenges in installing devices like security cameras or smart door locks. As P7 mentioned, *"Plus the costs, I mean, you have to consider how much it costs to, to do it. Um, and if say the lock gets broken or whatever gets broken or smashed, you know, it's more expensive to replace."*

4.2 Hosts' management of smart home devices

Hosts have access to data generated by guests with SHDs in STRs. We found hosts' lack of care accessing data, retention of data, and sharing data with SHDs. Hosts generally do not share data access with guests, but may do so with others (e.g., property managers); review data on their discretion; rarely delete data. Furthermore, our participants' data management falls short from their intentions to protect guests' privacy; they either are unaware or nonchalant about the privacy and security implications of their devices and data.

4.2.1 Who has access to SHDs' data?

Using hosts' accounts for SHDs in STRs. In general, SHDs need an account to access and control the devices and data. About half of our participants explicitly mentioned that they use their own accounts for their SHDs, which makes them the only ones who can access data collected by the SHDs. One common reason for using their own accounts is to provide a seamless experience for the guests. For example, P11 stated, *"[I use] my account. It's already very hard. Like people are traveling. It is already hard for them to kind of notice this. So as to make these changes. So I just logged them into my account."* Another reason is the complexity and trouble involved in adding a new user and removing them later. As stated by P14, *"you can adjust that [a smart home device] with your phone, but the guests don't have that on their phone... It's hard for the guests to have access to the smart devices 'cause they don't have my phone."*

Privacy and security implications when using guests' accounts for SHDs. Some of our participants reported that guests use their own accounts for the smart TVs and leave their accounts after they leave. This can result in privacy and security breaches, for example, hosts changing guests' account settings or viewing guests' browsing history [36]. Some participants admitted that they do not check if the accounts are logged in with guests' accounts. When this indeed happened, our participants reset or logged out of guests' accounts from the devices. P6 stated, *"we definitely actually reset them [smart TVs] for the next guest every time, because most of the time people forget to reset, like get out of the smart device."* However, P6 also mentioned that most guests do not care to ask to log out of their accounts.

Hosts may share data access with other stakeholders. Several participants mentioned various people who help them with their rental property, including cleaners, caretakers, property managers, and even neighbors. A few granted access/control of SHDs to their property managers to perform their duties more conveniently. For example, P2 stated, *"Yeah, [I will] give her full power over the thermostats because if she forgets to turn it down, and she goes home, she could just do it then."* P10, on the other hand, is a co-host and helps manage someone else's property. She mentioned that she has full access to the property owner's SHDs: *"I have complete access to everything, all of her, her passwords, etc., to work all of the smart devices."* These data management practices will be elaborated further in (section 4.2.3).

4.2.2 Managing control

Our participants wanted to control their SHDs in a way that respects guests' autonomy. Most participants only control their devices before and/or after guests' stay to take care of their properties (e.g., adjusting temperature).

Our participants noted that *"guest mode"* was an effective way to provide guests with the ability to control SHDs. For

example, almost all smart door lock owners created a guest-specific passcode to enter the property. P10 stated, *"I will also provide them with a code for the apartment door. And that I would send that to them, uh, two days before."* A few participants had smart TVs that also supported this feature, allowing hosts to set a guest profile. For devices that do not support guest modes, a few expressed concern about guests' accessing their information or abusing the access, a similar worry was reported in [15]. P7 stated, *"they [guests] could accidentally or intentionally do some, or like, 'Alexa, what are my last five orders?'...and get some kind information outta you."*

Although most participants wanted to make the guests comfortable, some of our participants wanted to have the ability to "override" guests' control. For example, P2 stated, *"if they [guests] turn the heat up to 90 degrees in the winter, I might be inclined to push it back down."*

4.2.3 Hosts' data management practices

In addition to who has access to the data, we were also interested in understanding how hosts manage the data collected by SHDs. We mainly considered two aspects of data management, which are reviewing and deleting data.

Reviewing data. Our participants monitored their smart cameras to make sure that guests have arrived and did not bring additional guests/pets. Unfortunately, unexpected guests/pets were a concern to many hosts, as they worry about insurance violation requirements, fire codes, or building's policies regarding guests. P11 voiced such frustration among many others, *"I don't know what for why, but for Airbnb, a lot of people organize parties, and that's not allowed in my Airbnb, nor in the condominium. It's forbidden. And so that's the reason why I have the cameras."* Some others also mentioned that they would check the camera to ensure the guests had left. P15 told us, *"And then time's like I know that the guest has checked out, so now that I can like begin cleaning."*

During guests' stay, our participants reported reviewing the data if they knew something went wrong. For example, P4 set up various notifications for events related to property damage (e.g., water sensor for flooding, sensors on oven for fire, unreasonable temperature settings). However, other participants relied on their gut feeling when monitoring guests. P11 stated, *"someone told me [he] didn't know how to use the espresso machine, which was a little weird because it was very simple...I didn't have a very good vibe about this guy. I saw him through the camera, and he was kind of pushing it like this. I'm like, oh my gosh, you're gonna break it."* Similarly, P10 mentioned he *"happened to check"* and noticed *"they [the guests] opened the door at some point, and then for like four or five hours, it wasn't closed. It wasn't locked."*

Deleting data. Our participants did not proactively delete data collected from their SHDs. A few reported that they did not delete video data collected from their smart cameras but instead relied on the devices' default expiration. For example, P1 stated, *"that is subject to Simplisafe's system. It gets deleted after 30 days or something."* P14, who owns a Blink camera, also mentioned something similar, but the duration of footage storage can be customized.

A few participants were reluctant to delete data because they needed the data for proof of business. P11 stated, *"you're supposed to tell Airbnb after 14 days, if any incident happened. So even if you try to delete, it's also not a good idea because I'm gonna be asked for stuff like that."* P13 shared a similar concern, stating, *"we don't delete anything that the camera's recording until after the stay. You know, until we know everything went well, the reviews are in all's well, and then we can delete everything."* We also found that few participants were unaware of smart speakers' data collection practices (e.g., access to conversation history).

4.3 Communicating smart home device usage with guests

Negotiating privacy starts with disclosure. Our participants knew and valued disclosing their devices, but also experienced its limitations; cameras were disclosed while other devices were neglected; hosts lacked accountability in disclosing their devices. Nevertheless, hosts viewed disclosure as an effective means to prevent and resolve future conflicts. Due to the lack of guideline, however, hosts' willingness to accommodate guests' privacy concerns were again left to their discretion.

4.3.1 Hosts' perceptions of platforms' guidelines

Both Airbnb and Vrbo have guidelines and policies regarding the use of smart devices on their properties. In summary, both platforms allow devices for security purposes, and only if they are disclosed beforehand. Airbnb, recently banned indoor cameras to respond to increasing concerns of hidden cameras [43]. Vrbo, on the other hand, does not allow any devices indoors unless they cannot be remotely controlled or disclosed, and guests can deactivate them [53]. Vrbo also has brief guidelines on managing data (e.g., limit access and deletion). Details of Airbnb/Vrbo's guidelines/policies regarding smart devices will be discussed in (section 5.3).

Hosts' familiarity with platforms' guidelines. To understand how our participants communicated their SHDs to guests, we first asked about their familiarity with platforms' guidelines/policies regarding the use of smart devices. Some of our participants were familiar with the guidelines/policies and were cognizant of them while setting up and editing their listing on the platform (e.g., through the prompted questions). P1 stated, *"I'm very familiar with them [the guidelines]. I*

knew about it...I think from setting up a new listing or editing an existing listing. Airbnb will notify you of fields that are incomplete, and I saw that, through the user interface that I showed you, the safety disclosures field that would allow me to add information about cameras."

Some of our participants who were not familiar with the guidelines/policies thought they were irrelevant to them. For example, P7 stated that since he does not have cameras on the property, *"never really had to look into it."* Similarly, P10 stated, *"we don't have any of the devices that they're talking about."* Too much focus on cameras and recording devices is problematic because, like P7 and P10, hosts can easily neglect to disclose other devices. Further, the definition of recording devices is ambiguous, as multi-function devices (e.g., devices with embedded audio/video) or interconnected devices can also potentially invade guests' privacy.

Notably, a few participants who were unfamiliar with the guidelines/policies thought that they were obvious. For example, P5 stated that although he did not know about the guidelines/policies, he *"kind of know[s] intuitively"*, mentioning that *"you don't want to have cameras inside."* Similarly, P8 stated that she was unfamiliar with the guidelines, but thought that *"the camera is like a requirement to be disclosed in an Airbnb listing"* and that *"if people don't do it, they're failing to follow the guidelines set by Airbnb and just common courtesy in general."* At the time of the study, Airbnb did not ban the use of indoor cameras, which means that there was a mismatch between host expectations and Airbnb requirements.

Hosts' perceptions about platforms' guidelines. About half of our participants were positive about their platforms' guidelines/policies. Our participants thought that they were concise (P7), understandable and reasonable (P13), and useful (P14). For example, P1 thought that it created a norm around the usage of smart devices on rental properties: *"I am glad that Airbnb provides the user interface to have specific disclosures of this because it gets people used to looking for that and, and it helps to create a norm for hosts to disclose any kind of surveillance devices."* Similarly, P4 thought it was a protection for both hosts and guests: *"I think it's a good expectation for renters, for guests to have, and it helps to keep hosts honest, because, you know, a lot of, I've rented from Airbnb, and it's like, half the time the information is wrong, they haven't filled out the stuff right, you know, you get messages from the hosts that are like, definitely for a different property...I think it's a protection both for guests and for hosts to have these sorts of policies from Airbnb."*

Some negative sentiments about the guidelines were that they were too generic (P14) or insufficient (P9). For example, P7 thought that the guidelines raised new questions (e.g., if smart doorbells would fit into any category). Similarly, some participants found the definitions unclear and confusing. For example, P9 was confused by the definition of common places (i.e., spaces without sleeping areas) that, *"I don't know if it's*

generalizable enough for every cases," and he preferred that Airbnb should not allow *"anything inside the house."*

Our participants have mixed opinions on what is considered a monitoring or surveillance device. They generally agreed that recording video or audio indoors invades guests' privacy. However, participants also mentioned that the intentions of the devices (P1) or the *"spirit of the device."* (P4) decide whether a device is considered as monitoring or as surveillance. For example, P1 distinguished between cameras and smart speakers in that cameras are generally security devices that are *"intentionally able to be used"* as surveillance devices because of *"how they can be used by the end user."* He added, *"the types of in-the-moment notice that is provided to users around the device whenever the device is listening or cameras are turned on,"* as a reason why smart speakers are not intended to be used to monitor or surveille people. On the other hand, P4 distinguished environmental monitoring (e.g., electricity, humidity) from surveillance, stating that *"the water meter is potentially a monitoring device,"* but *"that's not the spirit of that."*

Hosts' needs/wants in platforms' guidelines. To improve the negative aspects of platforms' guidelines/policies, our participants suggested that platforms need to increase both hosts' and guests' awareness about the existence of smart devices on property.

One way of increasing awareness is educating both hosts and guests to disclose and check whether or not there are smart devices on the property. P8 stated *"a little bit of education"* for both hosts and guests about *"how it [having smart home devices] benefits me [the host] other than just you know, me trying to creep on you [the guest]."*

Another way was to make it mandatory for hosts and guests to disclose and check the devices on the property. P1 pointed out that it is optional for hosts to disclose their devices. *"They [Airbnb] provide[s] that field among all the other fields that they provide when you're filling out a listing description."* Similarly, P4 commented how *"they [Airbnb] could probably do a little bit better job with helping hosts to implement that and to actually put it in front of guests' faces a little bit better."*

Circling back to P1's comment about how disclosing smart devices creates a norm around smart device usage in rental properties, our participants emphasized that this would be a combined effort from the hosts and the guests. P7 delivers this point: *"ultimately I imagine this is not really gonna be a legal thing either. It's just gonna be like a court of public perception. Like if customers demand that this be declared and disclosed, then it will be. And if they don't, then it won't. If enough people stop using Airbnb because people have Google homes, then Airbnb will require hosts to start declaring whether you have something or any kind of listening device."* (P7)

Some participants also suggested design mechanisms within the platform that could create more friction to increase awareness. P4 suggested a *"periodic checking"* from the plat-

forms, considering hosts might add devices after creating their listing. Similarly, P6 wanted notifications from the platform stating *"these are the important stuff. Send this notification to the guest before they arrive."* Furthermore, P16 suggested indicators for hosts (e.g., checkboxes, an asterisk on profile) and filters for guests to look for further information.

4.3.2 Hosts' communication of their SHDs to guests

Disclosing cameras is perceived as necessary.   About half of our participants thought that disclosing cameras on the property was necessary. P11 mentioned, *"I think absolutely [disclose] the cameras, because if you don't say about the cameras, you're gonna get in trouble."*

A few participants thought it was crucial to think of disclosing if cameras were located indoors or in private spaces (e.g., bathroom) because they believed that cameras can be used to monitor or surveille guests. For example, P1 stated, *"if they [cameras] were in the guest space, they certainly would need to be disclosed. That could potentially violate Airbnb's policy if it was in a private space, which I guess bathrooms, sleeping areas ... I think the fact that it's able to be manually used as a surveillance device, by which I mean, I can, even though I'm saying it's only turning on if the door is open, I can go into Simplisafe at any time and look through the cameras and record audio and video. So, I think that makes it important to disclose that it's visible, or that it's there."*

In terms of monitoring or surveillance, some participants mentioned that whether or not the devices had recording capabilities was an important factor to consider when disclosing devices. P7 stated, for example, *"I think if it's like recording someone, it would probably be good to notify people."*

Other devices are perceived as unnecessary to disclose. Other than cameras, our participants were unsure or thought it was unnecessary to disclose their smart devices. A few of our participants thought it was unnecessary to disclose smart speakers, which are considered privacy-invasive by guests [36, 54]. P7 mentioned, *"I don't think I would tell anyone that there was an Alexa dot or something in the listing. Both because someone might just go to your home and steal it, but also because, um, they don't really need to know. They could just unplug it if there's a problem."* Similarly, P4 stated that he has a Sonos soundbar with an embedded Google microphone in the living room, but *"don't feel that is a surveillance or monitoring device that would need to be disclosed."*

How hosts disclose their devices.   About half of our participants disclosed their smart devices on their rental profiles (e.g., listings, descriptions). Perhaps, the popularity to disclose on their rental profiles was because it was prompted as a default setting when hosts were listing their properties. For example, P1 stated that *"the only steps that we've taken is to use Airbnb built-in disclosure, to disclose the presence*

of a camera in the home." These built-in, default disclosures easily provided the hosts to check off the list of devices they have in the property. For example, P14 stated, *"there's a checkbox. Do you have these security devices and devices that are recording? If you check yes, then describe in detail, where is it located?"*

Some participants disclosed their devices at multiple points to ensure guests checked before booking or visiting the place. For example, in addition to disclosing their devices on their profiles, P4 disclosed his devices in his check-in instructions: *"the most important one is the one that we send before people [book]...[the message includes] there are cameras that are facing the two exterior doors."* Similarly, P6 disclosed her devices in the listing and a physical manual *"to make sure they're [guests] gonna do it [read or follow instructions]"* but also acknowledging that *"but most of the time, they're not gonna do it."* P11 shared a similar frustration after disclosing her devices in multiple points, stating *"It's not only on one point, it's on two. If people don't read it, they really need to get their act together, 'cause, I'm already disclosing it twice."*

4.3.3 Hosts' willingness to negotiate with guests

Conflicts with guests around SHDs usage.   A few of our participants experienced conflicts with guests around the usage of smart cameras. For example, P5 stated that one of his guests "looked at it [an outdoor smart camera] in an annoying way and then they stopped working". P5 did not respond to it because the guest was considered as a friend. P11, on the other hand, had several disputes with her guests about her cameras. Some guests were upset about her cameras in the living room and tampered them, which P11 reported to Airbnb as property damage. After receiving complaints from her neighbors about her guests making noises late at night, P11 monitored her guests through her cameras in the living room. The guests left a review of her being a predator, and P11 contacted Airbnb to remove those comments. A guest was taken aback when P11 warned the guest leaving trash outdoors with a photo taken from her outdoor camera. At the time of the study, Airbnb allowed cameras indoors, and in her defense, P11 disclosed them. These anecdotes support previous studies' finding about guests' discomfort in using monitoring devices [36, 54], however, also point out that for specific devices (e.g., smart cameras), disclosing might not be sufficient to mitigate guests' concerns.

Most of our participants, however, did not have conflicts around SHDs usage with guests (even for those who had smart cameras). However, they could still anticipate such situations. As a preventative measure, our participants expected platforms to provide a mechanism that ensures hosts disclose their devices and guests to read hosts' disclosure. P8 stated *"they [Airbnb] could potentially make their policy around camera and recording devices a little bit more clear for hosts. I think that some hosts are not, you know, they're*

not super tech savvy. They may not even realize that Amazon Echo records you, so they may not know that's something that should be disclosed," on the caveat that *"some people just may still not care or may not read them."* Some participants, therefore, suggested a mechanism to make sure that guests read the disclosure. P11 stated, *"I think for them [guests] signing a disclosure would be an extra step, and maybe not necessary, but I don't know what else. I mean, they're already consent[ing] by reserving."*

About half of our participants prefer direct communication with the guests if conflicts arise, even with preventative measures. P1 stated, *"I would first expect them [guests] to talk to me about it."* A few of our participants explicitly preferred messaging on the platform because it left them evidence. P6 stated, *"most of the time they [guests] send message to Airbnb and we answer them. This is because if something's happened first of all, Airbnb knows everything. And if they['re] trying to prove anything, you know."*

Hosts' willingness to accommodate guests' concerns   Our participants' willingness to accommodate guests' concerns was highly contextual; hosts considered factors such as duration of stay, trust, reasoning, and consequences.

First, a few participants mentioned that the length of the stay or the trust they built with the guests would matter. Contemplating whether he would disable his cameras, P1 stated, *"it would depend on what the guests' concern was related to and the level of trust we'd established with the person, since they are generally long-term roommates. I think if we had a conversation about those concerns, we would consider disabling the security feature, the cameras for security feature in order to help them feel more private and all that."* Similarly, P12 stated, *"if it's a longer stay, I would definitely say we will, we can turn the cameras off at the door and just keep the cameras on at the driveway so we can just monitor who's coming and who's going."*

A few participants also mentioned that they would first seek out the reasons for the requests. P14 stated, *"the security is not just for the guest but it's for myself personally, it's like my home, so I would need to know the reason why first."* A few participants were willing to negotiate, but with the caveat that the guests would be responsible for the consequences. P11 mentioned that *"we can disable the internal camera. I just want you to know that if there's a party and if I get a fee, you're gonna pay the full fee."*

About half of our participants were willing to accommodate guests' privacy concerns by disabling their devices. A few participants were willing to disable their cameras. For example, P1 stated, *"I think we would, depending on what, how that conversation went, I think I would consider disabling it [smart cameras]."*

A few participants stated that they would disable their smart speakers. P9 stated, *"I [will] definitely remove the voice assistant and for the future."* A few participants were willing

to disable other devices such as smart TVs, smart outlets, and lights. P7 stated, *"I have a couple smart outlets and um, you can, I can turn the outlets on or off from anywhere so I could turn 'em off and then I don't have to worry about it. Smart lights, you could probably turn to some kind of a dummy mode where they just work like regular lights."*

On the other hand, our participants also had devices that were non-negotiable for various reasons. For a few participants, cameras were a non-negotiable because they were worried about safety and security. P15 stated, *"I think when it comes to like a camera in a public space, it's like, well, what are you planning to do in that public space, let alone in the private space where I don't have a camera...that starts to get really fishy for me."* Other reasons include door locks being necessary to let guests in (P5) and when smart devices are installed and cannot be removed (P11).

5   Discussion

5.1   Concerning usage of smart cameras

Smart cameras were among the top three SHDs that our participants used. Our participants placed these cameras in private, shared, and public places to monitor their properties and guests to ensure safety and security. The ways in which our participants used smart cameras are concerning in many ways, especially considering that guests find the usage of cameras particularly privacy-invading [36, 54].

Our participants' reasons for using smart cameras were to monitor their properties, but by doing so they inevitably, if not intentionally, monitored their guests. For example, our participants place their smart cameras mostly in public spaces (e.g., front door) to count the number of guests arriving, to make sure that guests are bringing the appropriate number of guests. Also, a few participants installed smart cameras in shared spaces after experiencing theft, damage, and other violations of house rules.

At the time of the study, Airbnb allowed indoor cameras, and we had one participant (P11) who placed a camera in a private space (e.g., living room) to monitor guests. Granted, P11 struggled with guests not respecting house rules (e.g., parties) and disclosed upfront to guests about the cameras indoors. However, considering that our participants repeatedly complained that even if they put the effort to disclose SHDs, guests do not care to check, it is unlikely that guests would have known. In addition, when managing devices and data is up to hosts' discretion, monitoring the property can easily creep into monitoring guests. It is indeed a step forward to protecting guests' privacy that Airbnb updated their guidelines to ban indoor cameras, however, there is no accountability to make sure that preexisting cameras are removed from indoors.

Furthermore, the emphasis on smart cameras only opens up new questions: what about multi-function SHDs that have embedded audio/video capabilities? What about interconnected

devices? We touch upon this issue in (section 5.3) when we think about ways to improve platforms' guidelines/policies.

5.2 Hosts' efforts to protect guests' privacy falls short of their intentions

Previous research reported that privacy is less of a concern for Airbnb hosts when using SHDs, and if they do have concerns, it is about their own privacy instead of guests' (e.g., guests accessing hosts' information through SHDs) [15]. Similarly, Mare et al. [36] did not find guests' privacy as one of hosts' concerns when using SHDs. Our findings suggest quite the opposite: Airbnb hosts consider guests' privacy when they decide where to place smart cameras (Section 4.1), making sure they log out or reset the devices from guests' accounts (Section 4.2.1), monitoring data only when necessary (Section 4.2.3), restraining from controlling the devices when guests are visiting (Section 4.2.2), and making sure to disclose cameras on their properties (Section 4.3.2). All of these behaviors were previously unreported.

However, hosts' effort to protect guests' privacy falls short of their intentions, especially when it comes to managing data (Section 4.2.3). We found that our participants often did not have a clear threat model, which makes them unaware of the privacy implications behind their practices. For example, some of our participants shared accounts with property managers (e.g., housekeepers, cleaners) but did not consider them as potential entities that could infringe on guests' privacy. In addition, real-time monitoring by our participants was common. Granted, our participants tried to monitor their property and guests only when necessary (e.g., through notifications) but also admitted that they felt the urge to check on their property based on their "gut feelings", and did so. Similarly, our participants gave guests limited control and wanted to override their controls if necessary (e.g., temperature settings). Not to mention that hosts do not, if rarely, delete data collected by the devices after guests' visit.

Perhaps the lack of hosts' privacy-protecting measures that we've identified stem from the conflicting need between wanting to protect host's property and wanting to protect guest's privacy. The lack of guidance from platforms on how to use smart devices in a way that protects guests' privacy, and hosts' precarious position as gig-workers [10, 14, 48], requires constant proof-of-business from either side at the expense of guests' privacy or the security of property. Next, we discuss what we can and should do in terms of providing a clear guideline for hosts and guests to create a safe and privacy-protecting way to use smart devices in rental properties.

5.3 Insufficient guidelines/policies to support hosts

The STR platforms that we investigated in this study are Airbnb and Vrbo, and they each have their own guidelines and policies about the usage of smart devices on their properties.

According to Airbnb's guidelines regarding "use and disclosure of security cameras, recording devices, noise decibel monitors, and smart home devices" monitoring devices (e.g., security cameras, recording devices) are banned indoors but allowed outdoors if hosts disclose in listing's description. Noise decibel monitors (i.e., devices that assess sound level but do not record audio) are allowed indoors and hosts must disclose its presence, but not the location. Hosts are encouraged but not required to disclose SHDs and hosts are encouraged to provide options to guests to disable or unplug them [43]. Vrbo's policies regarding "surveillance devices at property" includes devices that capture image, audio, video, geolocation, personally identifiable information (PII), and internet activity. These devices are not allowed in the property, except for smart devices that cannot be remotely controlled if they are disclosed and given the option to disable them. Security cameras and smart doorbells may be used outside only for security purposes, if their locations are disclosed in multiple channels, access to data is limited, and deletion of data when no longer needed. Noise monitoring devices should be disclosed. These policies are enforced [53].

Although our intention was not to directly compare Airbnb's and Vrbo's guidelines and policies, we found that Vrbo's policies were clearer (e.g., definition) and more comprehensive (e.g., data management). That being said, it is important to note that not all participants were aware of their platforms' policies and guidelines, and this was true even for Vrbo, which had more rigorous policies than Airbnb's. Part of the problem is that some hosts are not aware of the existence of these policies and guidelines, and were not enforced to do so in the case of Airbnb. Even hosts who knew about the guidelines skimmed through it, thought it was irrelevant because they did not have security cameras/recording devices, or skipped disclosing other devices than security cameras/recording devices. The lack of guidance and enforcement to communicate whether hosts have smart devices in their properties is concerning and regrettable, considering how our participants thought these policies and guidelines were positive (e.g., creating a norm around smart device usage in rental properties) and acknowledged that it was a shared effort with involved stakeholders (e.g., matter of public opinion).

Therefore, we suggest and argue that platforms improve their policies and guidelines, not only with their contents, but also in how they address and enforce them with their users (e.g., hosts and guests). Many of the improvements to Airbnb's guidelines can be referred from Vrbo's policies, and we think this is important considering the prevalence and popularity of Airbnb for STRs [49].

Providing guidelines beyond types of devices. Our participants were confused with Airbnb's guidelines for devices other than cameras. For example, our participants were confused about smart speakers because they thought that smart

speakers do not intentionally record people and, therefore, are not recording devices. Airbnb has a more relaxed approach to SHDs; they "encourage" hosts to disclose SHDs and to give guests the option to disable or unplug the devices. Vrbo's definition of surveillance devices (i.e., devices that capture image, audio, video, geolocation, PII, internet activity) practically bans SHDs inside the property. The only exception is when these devices are not remotely controlled, which is difficult to achieve unless the host gives guests complete access/control of the devices. The ambiguity and complexity of platforms' guidelines/policies left our participants confused.

This is especially concerning since previous research found that Airbnb guests were concerned about devices that could potentially monitor them [36, 54], not to mention people's privacy concerns with smart speakers in general [1, 23, 29]. Therefore, we recommend that platforms consider the types of data collected by the devices, rather than the devices themselves when defining monitoring/surveillance devices.

Considering interconnected, multi-functioning devices. All of our participants agreed that smart cameras in bedrooms or bathrooms were unacceptable. However, at the time of the study, Airbnb allowed smart cameras indoors, and we had participants who placed smart cameras in private and shared areas (e.g., living room). Now, Airbnb bans smart cameras indoors. This shows that guidelines and policies regarding smart devices in STRs are evolving and ever-changing. Although Airbnb's recent update in its guidelines is a step towards protecting guests' privacy, there is more to consider. Hosts are still not required to disclose their SHDs, which is concerning when we think about SHDs with multiple functionalities (e.g., smart speakers with embedded cameras, smart TVs with audio/video capabilities). Furthermore, the interconnectivity of SHDs complicates the privacy implications (e.g., data shared among SHDs). As guidelines/policies shape norms in using SHDs in STRs, there is much more room for improvement.

Supporting SHDs' data management practices. Platforms should also provide guidelines/policies on how to manage data collected by SHDs because, ultimately, it is data privacy that matters to people's privacy concerns [51]. Currently, only Vrbo has some instructions for data management (e.g., limiting access to data, unnecessary data deletion), and Airbnb, as a more popular platform, should adopt these instructions in its guidelines. Account sharing and retention periods are other considerations in improving platforms' policies/guidelines. Considering that our participants shared their accounts with guests and property managers (e.g., housekeepers and cleaners), platforms should remind hosts to think about who they are sharing data with when they are sharing accounts. Next, since our participants depended on devices' default data deletion or were reluctant to delete data because they needed proof of business, platforms can suggest hosts delete data after guests' complaint period (e.g., 60 days).

Engaging both hosts and guests on SHDs disclosure. Disclosure is a form of *notice-and-choice*, which is prone to fail due to fatigue or negligence [47]. Thus, platforms should be more active during the disclosure process through better design, encouraging more engagement from hosts and guests.

For hosts, platforms should encourage, if not mandate, the disclosure of hosts' devices throughout their hosting and booking services, instead of just the beginning. Designers can make it mandatory for hosts to disclose their smart devices when registering their homes. The platform should also regularly check in with hosts to update their disclosures, considering hosts might add or remove their devices later. For guests, platforms should display the disclosure information more prominently on the listing, and should actively seek guests' confirmation by displaying such information during the guest's initial inquiry or the final confirmation of booking.

Currently, disclosure of security cameras is buried under irrelevant sections (e.g., the amenity section on Airbnb). When the SHDs in question is not a camera, the device does not even have its own dedicated space to be disclosed. Given that guests often prioritize price and location when booking for STRs, we believe asking for proactive consent from the guests could be more effective than passively showing such information on the listing only.

6   Conclusion

Our study explored the possibility of privacy negotiation between short-term rental (STR) hosts and guests by investigating hosts' practices on usage, management, and disclosure of smart home devices (SHDs), which provides valuable insight into hosts' willingness to accommodate guests' privacy. We conducted online interviews with 15 STR hosts and found that hosts consider guests' privacy when using SHDs: what types of devices they choose to use and locate them, logging out from guests' accounts, limiting monitoring, and disclosing cameras. However, we also found that hosts experience a dilemma between protecting their property versus protecting guests' privacy, therefore, making their efforts fall short of valuing guests' privacy. We identify platforms' insufficient support as the fundamental problem, leaving hosts astray in communicating with guests. We discuss ways to engage both hosts and guests to care about this matter by suggesting improvements to platforms' policies and guidelines, as well as design recommendations for features and functions to support communication.

Acknowledgments

References

[1] Noura Abdi, Xiao Zhan, Kopo M. Ramokapane, and Jose Such. Privacy Norms for Smart Home Personal Assistants. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14, May 2021.

[2] Amazon. How does drop in work? `https://www.amazon.com/gp/help/customer/display.html?nodeId=GS3WRTSRKD2U6MCK`.

[3] Amazon.com. Immersive voice experiences. `https://developer.amazon.com/en-US/alexa/alexa-for-hospitality`.

[4] Noah Apthorpe, Yan Shvartzshnaider, Arunesh Mathur, Dillon Reisman, and Nick Feamster. Discovering Smart Home Internet of Things Privacy Norms Using Contextual Integrity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(2):1–23, July 2018.

[5] Natã M. Barbosa, Joon S. Park, Yaxing Yao, and Yang Wang. "What if?" Predicting Individual Users' Smart Home Privacy Preferences and Their Changes. *Proceedings on Privacy Enhancing Technologies*, 2019(4):211–231, October 2019.

[6] Nicole Bergen and Ronald Labonté. "everything is perfect, and we have no problems": detecting and limiting social desirability bias in qualitative research. *Qualitative health research*, 30(5):783–792, 2020.

[7] Julia Bernd, Ruba Abu-Salma, Junghyun Choy, and Alisa Frik. Balancing power dynamics in smart homes: nannies' perspectives on how cameras reflect and affect relationships. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 687–706, 2022.

[8] Kelly Caine. Local standards for sample size at chi. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 981–992, 2016.

[9] Dan Calacci, Jeffrey J. Shen, and Alex Pentland. The cop in your neighbor's doorbell: Amazon ring and the spread of participatory mass surveillance. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2), nov 2022.

[10] Mingming Cheng and Carmel Foley. Algorithmic management: The case of airbnb. *International Journal of Hospitality Management*, 83:33–36, 2019.

[11] Yushi Cheng, Xiaoyu Ji, Tianyang Lu, and Wenyuan Xu. On detecting hidden wireless cameras: A traffic pattern-based approach. *IEEE Transactions on Mobile Computing*, 19(4):907–921, 2019.

[12] CNET. 9 devices every airbnb host should put in their rental, 2018. `https://www.cnet.com/home/smart-home/devices-every-airbnb-host-should-put-in-their-house/`.

[13] Camille Cobb, Sruti Bhagavatula, Kalil Anderson Garrett, Alison Hoffman, Varun Rao, and Lujo Bauer. "I would have to evaluate their objections": Privacy tensions between smart home device owners and incidental users. *Proc. Priv. Enhancing Technol.*, 2021(4):54–75, 2021.

[14] Suzanne C De Janasz, Sowon Kim, Joy Schneer, Nicholas J Beutell, and Carol Wong. Work-family integration and segmentation in the gig economy: Airbnb hosts' challenges and strategies. In *Academy of Management Proceedings*, volume 2020, page 20130. Academy of Management Briarcliff Manor, NY 10510, 2020.

[15] Rajib Dey, Sayma Sultana, Afsaneh Razi, and Pamela J. Wisniewski. Exploring Smart Home Device Use by Airbnb Hosts. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, pages 1–8. Association for Computing Machinery, 2020.

[16] Pardis Emami-Naeini, Henry Dixon, Yuvraj Agarwal, and Lorrie Faith Cranor. Exploring How Privacy and Security Factor into IoT Device Purchase Behavior. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, May 2019.

[17] Rob Gabriele. Vacation surveillance: What travelers think about airbnb security cameras, 2023. `https://www.safehome.org/home-security-cameras/traveler-perceptions-airbnb-cameras/`.

[18] Radhika Garg and Christopher Moreno. Understanding Motivators, Constraints, and Practices of Sharing Internet of Things. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(2):1–21, June 2019.

[19] Tomas Gecevicius, Yaliang Chuang, and Jingrui An. Smart arbnb: Smart home interface for airbnb with augmented reality and visible light communication. In *CHIIoT@ EWSN/EICS*, 2021.

[20] Christine Geeng and Franziska Roesner. Who's in control? interactions in multi-user smart homes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019.

[21] Yangyang Gu, Jing Chen, Cong Wu, Kun He, Ziming Zhao, and Ruiying Du. Loccams: An efficient and robust approach for detecting and localizing hidden wireless cameras via commodity devices. *Proceedings of the*

ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 7(4):1–24, 2024.

[22] Weijia He, Nathan Reitinger, Atheer Almogbil, Yi-Shyuan Chiang, Timothy J. Pierson, and David Kotz. Contextualizing interpersonal data sharing in smart homes. *Proceedings on Privacy Enhancing Technologies*, 2024(2), 2024.

[23] Yue Huang, Borke Obada-Obieh, and Konstantin Beznosov. Amazon vs. my brother: How users of shared smart speakers perceive and cope with privacy risks. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–13, 2020.

[24] Md Nazmul Islam and Sandip Kundu. Preserving iot privacy in sharing economy via smart contract. In *2018 IEEE/ACM Third International Conference on Internet-of-Things Design and Implementation (IoTDI)*, pages 296–297. IEEE, 2018.

[25] William Jang, Adil Chhabra, and Aarathi Prasad. Enabling multi-user controls in smart home devices. In *Proceedings of the 2017 workshop on internet of things security and privacy*, pages 49–54, 2017.

[26] Haojian Jin, Boyuan Guo, Rituparna Roychoudhury, Yaxing Yao, Swarun Kumar, Yuvraj Agarwal, and Jason I. Hong. Exploring the Needs of Users for Supporting Privacy-Protective Behaviors in Smart Homes. In *CHI Conference on Human Factors in Computing Systems*, pages 1–19, April 2022.

[27] Jungsun Kim, Mehmet Erdem, and Boran Kim. Hi alexa, do hotel guests have privacy concerns with you?: A cross-cultural study. *Journal of Hospitality Marketing & Management*, pages 1–24, 2023.

[28] Vinay Koshy, Joon Sung Sung Park, Ti-Chung Cheng, and Karrie Karahalios. "we just use what they give us": Understanding passenger user perspectives in smart homes. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2021.

[29] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. Alexa, are you listening? privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proceedings of the ACM on human-computer interaction*, 2(CSCW):1–31, 2018.

[30] Tu Le, Alan Wang, Yaxing Yao, Yuanyuan Feng, Arsalan Heydarian, Norman Sadeh, and Yuan Tian. Exploring smart commercial building occupants' perceptions and notification preferences of internet of things data collection in the united states. In *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*, pages 1030–1046. IEEE, 2023.

[31] Tu Le, Zixin Wang, Danny Yuxing Huang, Yaxing Yao, and Yuan Tian. Towards real-time voice interaction data collection monitoring and ambient light privacy notification for voice-controlled services.

[32] Patricia Leavy. *Research design: Quantitative, qualitative, mixed methods, arts-based, and community-based participatory research approaches*. Guilford Publications, 2022.

[33] Anna Lenhart, Sunyup Park, Michael Zimmer, and Jessica Vitak. "You Shouldn't Need to Share Your Data": Perceived Privacy Risks and Mitigation Strategies Among Privacy-Conscious Smart Home Power Users. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):247:1–247:34, October 2023.

[34] Heather Richter Lipford, Madiha Tabassum, Paritosh Bahirat, Yaxing Yao, and Bart P Knijnenburg. Privacy and the internet of things., 2022.

[35] Diba Malekpour Koupaei and Kristen Cetin. Smart thermostats in rental housing units: Perspectives from landlords and tenants. *Journal of Architectural Engineering*, 27(4):04021042, 2021.

[36] Shrirang Mare, Franziska Roesner, and Tadayoshi Kohno. Smart devices in airbnbs: Considering privacy and security for both guests and hosts. *Proc. Priv. Enhancing Technol.*, 2020(2):436–458, 2020.

[37] Karola Marky, Nina Gerber, Michelle Gabriela Pelzer, Mohamed Khamis, and Max Mühlhäuser. "you offer privacy like you offer tea": Investigating mechanisms for improving guest privacy in iot-equipped households. *Proceedings on Privacy Enhancing Technologies*, 2022.

[38] Karola Marky, Alexandra Voit, Alina Stöver, Kai Kunze, Svenja Schröder, and Max Mühlhäuser. "I don't know how to protect myself": Understanding Privacy Perceptions Resulting from the Presence of Bystanders in Smart Environments. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*, pages 1–11, October 2020.

[39] Dana McKay and Charlynn Miller. Standing in the Way of Control: A Call to Action to Prevent Abuse through Better Design of Smart Technologies. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pages 1–14, New York, NY, USA, May 2021. Association for Computing Machinery.

[40] Phoebe Moh, Pubali Datta, Noel Warford, Adam Bates, Nathan Malkin, and Michelle L. Mazurek. Characterizing Everyday Misuse of Smart Home Devices. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 2835–2849, May 2023.

[41] Sara Morrison. Google assistant's new guest mode is more private, but there's a trade-off, 2021. https://www.vox.com/recode/22229008/google-assistant-guest-mode.

[42] Savvas Papagiannidis and Dinara Davlembayeva. Bringing smart home technology to peer-to-peer accommodation: Exploring the drivers of intention to stay in smart accommodation. *Information systems frontiers*, 24(4):1189–1208, 2022.

[43] Community policy. Use and disclosure of security cameras, recording devices, noise decibel monitors, and smart home devices. https://www.airbnb.com/help/article/3061.

[44] Johnny Saldaña. *The Coding Manual for Qualitative Researchers*. SAGE, 2nd ed edition, 2013.

[45] Shane Schutte. *Tenant Sentiment Effects of Smart Home Technology in Short-Term Rentals*. PhD thesis, University of Nebraska at Omaha, 2023.

[46] skatun. Smart speakers treated as surveillance. https://airhostsforum.com/t/smart-speakers-treated-as-surveillance/31758.

[47] Robert H Sloan and Richard Warner. Beyond notice and choice: Privacy, norms, and consent. *J. High Tech. L.*, 14:370, 2014.

[48] Robert Sprague. Are airbnb hosts employees misclassified as independent contractors? *U. Louisville L. Rev.*, 59:63, 2020.

[49] Statista. Airbnb - statistics & facts. https://www.statista.com/topics/2273/airbnb/#topicOverview.

[50] Aron Szanto and Neel Mehta. A host of troubles: Re-identifying airbnb hosts using public data. *Technology Science. Oct*, 2018.

[51] Madiha Tabassum, Tomasz Kosinski, and Heather Richter Lipford. "I don't own the data": End user perceptions of smart home device data practices and risks. In *Fifteenth symposium on usable privacy and security (SOUPS 2019)*, pages 435–450, 2019.

[52] Parth Kirankumar Thakkar, Shijing He, Shiyu Xu, Danny Yuxing Huang, and Yaxing Yao. "it would probably turn into a social faux-pas": Users' and bystanders' preferences of privacy awareness mechanisms in smart homes. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2022.

[53] Vrbo. Vrbo's policy on surveillance devices at a property. https://help.vrbo.com/articles/What-is-HomeAway-s-policy-on-surveillance-devices-at-a-property.

[54] Zixin Wang, Danny Yuxing Huang, and Yaxing Yao. Exploring Tenants' Preferences of Privacy Negotiation in Airbnb. In *Proceedings of the 32nd USENIX Security Symposium*, 2023.

[55] Chris Welch. The wynn las vegas is putting an amazon echo in every hotel room, 2016. https://www.theverge.com/circuitbreaker/2016/12/14/13955878/wynn-las-vegas-amazon-echo-hotel-room-privacy.

[56] Meredydd Williams, Jason R C Nurse, and Sadie Creese. "Privacy is the Boring Bit": User Perceptions and Behaviour in the Internet-of-Things. In *15th International Conference on Privacy, Security and Trust (PST)*, 2017.

[57] New York Times Wirecutter. If you've turned your home into an airbnb, you need smart devices, 2020. https://www.nytimes.com/wirecutter/blog/airbnb-smart-devices/.

[58] Kevin Wu and Brent Lagesse. Do You See What I See? Detecting Hidden Streaming Cameras Through Similarity of Simultaneous Observation. In *2019 IEEE International Conference on Pervasive Computing and Communications*, pages 1–10, March 2019.

[59] Yaxing Yao, Justin Reed Basdeo, Smirity Kaushik, and Yang Wang. Defending my castle: A co-design study of privacy mechanisms for smart homes. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–12, 2019.

[60] Yaxing Yao, Justin Reed Basdeo, Oriana Rosata Mcdonough, and Yang Wang. Privacy perceptions and designs of bystanders in smart homes. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–24, 2019.

[61] Eric Zeng, Shrirang Mare, Franziska Roesner, Santa Clara, Eric Zeng, Shrirang Mare, and Franziska Roesner. End User Security and Privacy Concerns with Smart Homes. In *Proceedings of the 13th Symposium on Usable Privacy and Security*, 2017.

[62] Shikun Zhang, Yuanyuan Feng, Yaxing Yao, Lorrie Faith Cranor, and Norman Sadeh. How usable are ios app privacy labels? *Proceedings on Privacy Enhancing Technologies*, 2022.

[63] Yunxia Zhu, Mingming Cheng, Jie Wang, Laikun Ma, and Ruochen Jiang. The construction of home feeling by airbnb guests in the sharing economy: A semantics

perspective. *Annals of Tourism Research*, 75:308–321, 2019.

[64] Yixin Zou, Kaiwen Sun, Tanisha Afnan, Ruba Abu-Salma, Robin Brewer, and Florian Schaub. Cross-Contextual Examination of Older Adults' Privacy Concerns, Behaviors, and Vulnerabilities. *Proceedings on Privacy Enhancing Technologies*, 2024.

A   Interview protocol

Hi, thanks for joining us today. I am [name] from [institution] and these are my colleagues. Today, we'd like to talk with you about your experiences using SHDs on your STR property. Before we begin, we kindly request your consent for this study.

To proceed with this process, I will start the recording now, and then we can go ahead with the oral consent. The recordings will only be accessed by us [two or three depending on who are in the meeting].

May I please have your name and today's date? Do you agree to participate in this study? Can you please verify the code that we've sent to your [Airbnb/Vrbo] account?

Thank you for your patience and understanding. We emphasize that you are under no obligation to answer any questions that you are uncomfortable with. You are free to skip any questions and you can withdraw from this study at any time you wish. Do you have any questions before we start? [Take questions] If you have no further questions about the study, let's start.

We are interested about your experiences with using SHDs on your STR property. First, we would like to define what SHDs are: SHDs are household items that are connected to the Internet or a home network to enhance functionality, connectivity, and efficiency within the home. Examples include smart speakers (e.g., Amazon Echo or Google Home), smart lights (e.g., Philips Hue), smart thermostats (Google Nest), and smart locks and security cameras. Please note that we do not include personal devices such as computers, smartphones, tablets, and smartwatches in our definition. Having that in mind, we first wanted to ask about your motivations on using SHDs on your property. [check with survey entry and ask] Can you describe your [Airbnb/Vrbo] property? (e.g., how many, what type, layout) What types of SHDs do you have on your [Airbnb/Vrbo] property? Why did you buy those devices? Where are your SHDs located and what was the reason for locating them there? prompt: How long have you had those SHDs on your [Airbnb/Vrbo] property?

Next, we want to know more about your experience of using SHDs on your STR property. How do you manage your SHDs when your guests visit? (e.g., changing devices' location, turning on/off the devices, reminding guests of their presence, managing accounts and data). Are there any things you do differently to manage your SHDs *before/during/after*

guests' visit? Are there any challenges you had using SHDs on your [Airbnb/Vrbo] property with your guests? Any conversations? What did you talk about? prompt: were there any guest comments that mentioned SHDs on your [Airbnb/Vrbo] property?

We're especially interested in how you disclose your devices to your guests. Can you tell us if and how you described your SHDs in your [Airbnb/Vrbo] listings? Are there any considerations you have when disclosing SHDs to your guests? (e.g., types of devices, location of devices) Are there SHDs you absolutely think you need to disclose to your guests? Are there SHDs you don't think you need to disclose to your guests? Why?

Thinking about how to disclose your SHDs to your guests, When would you like to do it, and how? (e.g., through the listing, when guests are booking, reminding guests through manual)

So, Airbnb allows the use of cameras and recording devices if they are disclosed in the public and common spaces. Airbnb does not allow if devices are not disclosed and/or are in public spaces. [show Airbnb guidelines; make sure to zoom in when sharing screen; can also share link in chat; make sure to give enough time for participants to read] [2] How familiar are you with the guidelines? How did you know about it? What do you think about the guideline? (e.g., understandable? useful?)

We're also curious about your concerns when using SHDes on your [Airbnb/Vrbo] property. What do you think are the benefits of having SHDs on your [Airbnb/Vrbo] property? What do you think are the drawbacks of having SHDs on your [Airbnb/Vrbo] property? Can you think of any potential issues that might arise from using SHDs on your [Airbnb/Vrbo] property? What are the issues that you might face as a host? What are the issues that guests might face?

[If participant does not mention any privacy issues] Here are some privacy issues that might arise from using SHDs in a STR property. For example, there was a debate among Airbnb hosts about a guest complaining that they were uncomfortable with smart speakers and whether hosts should disclose and/or use them [46]. Have you experienced any of these issues? What happened? How did you resolve these issues? If not, do you think any of these issues might happen to you? What would you do?

Thinking of the people involved in resolving these issues, (If that happened) What did your guests do? (If not) What might you want guests to do to resolve those conflicts? What might you want [Airbnb/Vrbo] to do to resolve those conflicts? What do you think might help to prevent these issues?

[check if team members have any remaining questions left] Thank you for your time today, before we end, we would like to ask if there was anything you wanted to say but didn't get to, or if you had any questions for us.

---

[2]We introduced Vrbo's policies for participants who hosted on Vrbo

B   Final codes, sub-codes, and their descriptions

Codes and sub-codes	Descriptions
STR property description	*Participants' description of their STRs regarding...*
Size	how big the property is in terms of the types of property hosted (e.g., entire house, private room) and the types of rooms hosted (e.g., 2 bed 2 bath).
Residence	the objectives of residence of the property hosted (e.g., primary, secondary, investment).
Characteristics	how the property is marketed to the guests (e.g., farmhouse) often shaped by local events and seasonality (e.g., ski-event, metro-area).
Guest characteristics	the types and sizes of guests (e.g., business, family).
Types and location of SHDs	*Participants mentioning what kind of devices they use and where.*
Motivations for using smart home devices	*Participants' description of why they use SHDs (similar to Advantages but different in that these are tied to intentions and expectations).*
Home automation	To automate their home.
Guest experience	To improve guests' experience (e.g., providing a seamless experience).
Monitoring property	To monitor the house.
Monitoring guests	To monitor guests (e.g., to check the number of guests).
Safety and security	For safety and security reasons both for the property and guests (e.g., fire, flooding).
Energy conservation	To save energy (e.g., temperature, lights).
Remote control	To have control of their property from remote.
Interconnectivity	To connect with other devices in the house.
Motivations for NOT using smart home devices	*Any reasons for participants' reluctance to using SHDs due to (similar to Disadvantages but different in that these are more tied to intentions and expectations)...*
Guests' privacy	Concerns about guests' privacy.
Cost	Concerns about cost.
Theft	Concerns about theft.
Technical difficulties	Concerns about the technical difficulties involved in using SHDs (e.g., installation).
Mindfulness	Wanting to provide guests with time away from technology.
No need	Participants' disinterest in using (specific) smart home devices.
Advantages of using smart home devices	*Participants' comment on the advantages of using smart home devices on their rental property (similar to Motivations but different in that these are lived experiences).*
Conserve energy	Using smart home devices saves energy consumption.
Safety and security	Using smart home devices provide safety and security to the property and guests.
Proof of business	Using smart home devices provide hosts with proof of business when they need evidence.
Entertainment	Using smart home devices provide entertainment for guests (e.g., smart TVs).
Visibility	Using smart home devices provide visibility (e.g., guests' activity, smart home data) to hosts.
Guest experience	Using smart home devices provide a better experience for guests (e.g., seamless entry).
Remote control	Using smart home devices provides the convenience of remotely controlling the property.

Codes and sub-codes	Descriptions
Disadvantages of using smart home devices	*Participants' experiences of challenges in using smart home devices on their rental property (similar to Motivations for NOT using SHDs but different in that these are more lived experiences).*
Technical failures (internal)	Facing technical failures due to device defects such as loss of network connection and/or battery outage.
Technical difficulties	Facing technical difficulties due to one's lack of technical proficiency, such as smart home devices being too complicated to manage.
(Potential) Theft	Guests stealing stuff from home, especially smart home devices.
Guest confusion	Guests being confused on how to use smart home devices on rental property.
Invasion of guests' privacy	Using smart home devices invades guests' privacy.
Guest misuse	Guests using smart home devices in a way that is not intended by the host (e.g., purchase of items)
Lack of guest control	Participants not being able to provide enough control for guests.
Reasons for device purchase and usage	*Reasons that participants buy certain devices and use it in a certain way.*
External sources	Participants learn and/or hear from external sources (e.g., other Airbnb hosts, online forums).
Integration	Choosing a specific brand/company to integrate the devices.
Brand reputation and trust	Preference based on reputation and trust to a specific brand/company.
Smart home device management	*Participants' management of the smart home devices and data.*
Accounts and passwords	Whose accounts are being used and how passwords are shared for the smart home devices.
Shortcuts	Creating workarounds to manage their smart home devices (e.g., using text shortcuts to remotely manage devices).
Access control	Access control mechanisms for guests using smart home devices (e.g., manual control, smartphone apps).
Notifications	Setting up notifications to manage SHDs.
Disclosure	Participants consideration of disclosure as a part of their management.
Routines and/or schedules	Participants set routines and schedules to manage their devices
Manual management	Physically managing the devices
Reviewing and deleting data	Reviewing and deleting data collected by SHDs at any time of the guests' stay.
Upgrade	Upgrading soft/hardware for SHDs
Disclosing smart home devices	*Participants' perception of whether or not to disclose any/certain SHDs to guests.*
Must	Smart home devices that participants think they absolutely should disclose to their guests.
Unsure	Smart home devices that participants are unsure if they should disclose to their guests.
Not disclose	Smart home devices that participants do not disclose to their guests for any reason.
Methods of disclosure	*How participants disclose/communicate their SHDs to guests (similar to Disclosure considerations but different in that these are actual practices).*
Property listing	Participants disclosing their SHDs in the property listings.
Property description	Participants disclosing their SHDs in the property descriptions.
Messaging	Participants disclosing their smart home devices through messaging (e.g., platform, external chats) with the guests.
Additional instructions	Participants disclosing their smart home devices with additional instructions (e.g., physical manual)
Preference for disclosure methods	*Participants' preference in how to disclose their SHDs.*

Codes and sub-codes	Descriptions
Disclosure considerations	*The kinds of considerations participants put into when they are thinking of disclosing SHDs to guests (similar to Methods of disclosure but different in a way that it might not be practiced).*
Instruction for guests	Leaving additional instructions for guests.
Respect to guest privacy	Thinking about how to respect guests' privacy when disclosing SHDs
Accounts	Disclosing who's account is associated with the device.
Data visibility	If participants consider disclosing what data is visible to whom.
Familiarity with STR policies/guidelines regarding SHDs	*Participants' self-reported familiarity with platforms' policies/guidelines regarding smart devices.*
Perceptions of platforms' policies/guidelines	*Participants' perceptions of platforms' policies/guidelines.*
Positive	Any positive reactions to platforms' policies/guidelines regarding smart devices.
Negative	Any negative reactions to platforms' policies/guidelines regarding smart devices.
Neutral	Any neutral reactions to platforms' policies/guidelines regarding smart devices.
Perception of surveillance/monotoring	*Participants' perceptions of surveillance versus monitoring.*
Needs/wants in platforms' policies/guidelines	*Participants' wants and needs regarding platforms' policies/guidelines regarding smart devices.*
Willingness to negotiate/accommodate	*Participants' willingness to negotiate, accommodate, and compromise with guests about their usage of SHDs to meditate guests' concerns.*
Non-negotiables	*Participants' reluctance to negotiate and reasons why.*
(Potential) conflicts with guests	*(Potential) conflicts with guests regarding the usage of SHDs in rental property.*
Resolving conflicts	*How participants resolve, or plan to resolve conflicts with guests regarding the usage of SHDs in rental property.*
Expectations towards guests	Participants' expectations of guests when they are in conflict.
Communication	Participants' mentioning of communication as a strategy to resolve conflicts with guests around the usage of SHDs.
Expectations towards platforms	Participants' expectations of platforms when they are in conflict with guests (e.g., moderation)
Transparency	An emphasis on being transparent about the usage of SHDs to guests when resolving conflicts.
Empathy	An emphasis on being empathetic to guests when resolving conflicts (e.g., If I were a guest...).
Explanation	An emphasis on explaining to guests the details of using SHDs with guests (e.g., why, where, how).
Granting access	Participants granting access to guests to resolve conflicts around SHDs.
Unreasonable requests from guests	*When participants think guests' requests are unreasonable, therefore no need to accommodate.*
Other stakeholders	*People involved in managing the rental property beside the hosts (e.g., caretakers, neighbors, cleaners, and property managers).*
Unsure	*Quotes that are interesting but unsure where they fit.*
Good quotes	*Quotes that represent, highlight codes and themes; Make sure that ends up in the writing.*
Tech-savvy host	*Participants who have indications that they are tech-savvy (e.g., background in IT).*

Codes and sub-codes	Descriptions
SHDs in the background	*Participants' strategies to deploy SHDs in a way that delivers guests with a seamless experience.*
Smart home adoption	*How participants adopt SHDs to their STRs (e.g., gradual).*
Needs/wants in SHD functionality	*Participants' needs/wants in SHD functionality to ease the use in their STRs (e.g., guest mode).*
Negotiation practices	*Any negotiation practices employed by participants to resolve conflicts with guests.*
Interesting but irrelevant	*Quotes that are interesting and potentially relevant to answering our RQs.*

Batman Hacked My Password:
A Subtitle-Based Analysis of Password Depiction in Movies

Maike M. Raphael [ID]
Leibniz University Hannover

Aikaterini Kanta [ID]
University of Portsmouth

Rico Seebonn [ID]
Leibniz University Hannover

Markus Dürmuth [ID]
Leibniz University Hannover

Camille Cobb [ID]
University of Illinois Urbana-Champaign

Abstract

Password security is and will likely remain an issue that non-experts have to deal with. It is therefore important that they understand the criteria of secure passwords and the characteristics of good password behavior. Related literature indicates that people often acquire knowledge from media such as movies, which influences their perceptions about cybersecurity including their mindset about passwords. We contribute a novel approach based on subtitles and an analysis of the depiction of passwords and password behavior in movies. We scanned subtitles of 97,709 movies from 1960 to 2022 for password appearance and analyzed resulting scenes from 2,851 movies using mixed methods to show what people could learn from watching movies. Selected films were viewed for an in-depth analysis.

Among other things, we find that passwords are often portrayed as weak and easy to guess, but there are different contexts of use with very strong passwords. Password hacking is frequently depicted as unrealistically powerful, potentially leading to a sense of helplessness and futility of security efforts. In contrast, password guessing is shown as quite realistic and with a lower (but still overestimated) success rate. There appears to be a lack of best practices as password managers and multi-factor authentication are practically non-existent.

1 Introduction

Cybersecurity is a topic that virtually everyone encounters every day, from the first unlocking of the smartphone in the morning to reading emails at work or communicating with friends at night. This requires many decisions about which links to click, which websites to trust or which password to choose. Among other things, these decisions are based on knowledge and beliefs about the subject area in question that determine, for example, what is perceived as "secure" or "insecure" [20, 73]. It is therefore important that this knowledge is correct and beliefs are aligned with reality. However, studies show that for cybersecurity these are often incorrect or incomplete, leading to "bad" security practices [1, 63].

This problem becomes particularly evident in password security, which is an area where many misconceptions are found. Various studies show that people do not know the characteristics of good passwords, do not know how to handle passwords in general or do not remember the recommendation to change the password regularly, which has been proven to be bad advice in recent years [11, 24, 45]. This is a big issue because, despite their weaknesses and numerous alternatives being available, passwords are still by far the most common form of online authentication [26]. It is therefore important that the understanding of password security and good password behavior is reinforced.

One source that influences people's perception of cybersecurity is likely films [54, 73]. Literature shows that people learn from media and use it as a source of information [51, 54]. Films play a major role in this; they have been a popular entertainment medium for decades and are watched by thousands of people every day. Therefore it is hardly surprising that films influence knowledge and behavior [19] or the attitude toward technology [9] and that this may change how people handle specific topics and make decisions. Because people often cannot decide if what they see is realistic or not, they are in danger of taking fictional portrayals as realistic which may influence their thinking about certain topics including cybersecurity [20, 73]. It is therefore important to ensure that things are presented in a good and realistic way so that people potentially learn something *true* from them [12].

The use of certain technologies in movies "both reflects and influences society's use and attitudes toward the portrayed

technology" [9]. Literature shows that what we see in the media (partially) reflects life [19] and shows what our society thinks and is interested in [21, 67]. Furthermore, movies can be regarded as "Cultural Artifacts" and historical snapshots" [9] enabling us to compare attitudes from different years. So we can use those to learn something about our society during the ages and to identify password behavior that seems to be considered typical as is already shown impressively for society's attitude toward technology.

Regarding the many misconceptions concerning passwords, it is to be expected that their portrayal in films may reflect an outdated or insecure password behavior that we also see in society. At the same time, showing good passwords and secure behavior could have a significant impact on the understanding and the overall security of people regarding password usage. Consequently, the aim of this paper is to answer the question of how accurately cybersecurity topics in general and password-related topics in special are depicted in films.

In this paper, there are four primary contributions to improve understanding of password depiction:

- A subtitle analysis as a novel approach to analyzing the occurrence of passwords within a large amount of data: 97,709 movies of various genres, from 1960 to 2022.
- A collection of which films and scenes and in which context passwords play a role and a statistic evaluation of the results from 2,851 movies. This includes what the password is used for, different kinds of password behavior and (missing) best practices such as the use of password managers and multi-factor authentication.
- An evaluation of the strength of the passwords shown in films and a linking of this knowledge with the results from step two to understand the role of *strong* passwords in movies.
- In particular, an investigation is conducted on attacks on passwords to understand whether passwords are presented as "secure" and which circumstances lead to blighting password protection. This includes watching selected movies to understand the overall importance and ambient conditions of password attack movie scenes.

The results show how many everyday password activities are mirrored in films and how often the topic appears in a wide variety of genres and years. However, this often involves insecure behavior such as careless sharing by writing down or reusing passwords for different accounts. Good password practices such as using password managers or multi-factor authentication is scarcely depicted in films. Even if strong passwords are used, these hardly increase security – passwords within the highest strength category (as shown in Section 6) are guessed as often and easily as very bad ones. Both hacking and guessing attacks are often frighteningly successful, which gives the impression that passwords can hardly withstand any attacks. However, there is a strong contrast between the very unrealistic hacking attacks and password guessing, which is often portrayed in a very realistic manner.

2 Background & Related Work

We first describe a small but closely-related body of work focused on movies and cybersecurity. We looked to the fields of film studies and linguistics to inform our understanding of methodological best practices and the way that movies impact people and society. Finally, since we focus on the depiction of passwords, existing knowledge about passwords, password security, and user experiences with passwords provided important context for structuring our analysis and interpreting our findings.

Cybersecurity and Movies   Prior work has shown that many users learn about cybersecurity from media, including advertisements, news, and fictional narratives such as television and movies [51, 52, 54].

Specifically, researchers have studied the impact of movies on people's understanding, perspectives, and behaviors related to hacking [4, 20], biometric and non-biometric authentication methods [73], and technology broadly [9], finding that movies have the capacity to misinform people or guide them toward a better understanding of technology. Authors argued that the movies sometimes confirmed participants' existing mental models, for example beliefs that only famous or rich people will be attacked and that – if targeted – security measures are futile anyway [20]. Other prior work helps us understand the mechanisms through which media such as movies might influence people [30]. For example, mental models are thought to be an important aspect of decision-making [30]; thus, researchers finding that watching movie scenes impacted mental models [4, 20] suggests that movies could influence people's decisions and behaviors as well. Perhaps the visual and/or video format of movies contributes to their ability to influence people; studying the difference in impact of a video-based message or a text-based message, Albayram et al. found that people who watched videos were more likely to adopt password managers [2]. Another influential factor may be the narrative structure that is common in fictional media such as movies. Prior work has repeatedly found that we learn about security through stories [46, 48].

Since movies have the capacity to (mis)inform, it is pertinent to understand their contents and to what extent this content is realistic. Examples poking fun at the inaccuracies of cybersecurity in movies are easy to find in blog articles [53], online repositories of TV tropes [16, 17], and even in a talk at DefCon [38]. These sources emphasize inaccuracies such as hacking or decryption being absurdly easy or quick, technical terms being thrown around without real meaning, and images of illuminated screens with rushing lines of code. Similarly, Christmann et al. find inaccuracies with password advice in YouTube videos and propose a list of requirements for security awareness videos dealing with password behavior [12]. In a more systematic study, Gordon assembled and analyzed a data set of 50 "hacker movies" from the 1960s

through the early 2000s, comparing the key themes in these movies with reality [23]. Gordon found that some aspects of movies' portrayals of hackers was quite realistic (e.g., finding that the inaccurate "stereotypical view of outsider attacks by teenagers" is *not* coming from this set of movies), while some were not (e.g., the ratio of insider to outsider attacks), but argues overall that these movies are likely to be a useful resource for security course instructors.

Learning From and With Film Media   Film media (i.e., television and movies) can have a positive influence on adults' or children's learning and influence them to adopt beneficial attitudes and behaviors [18, 34, 35, 65, 68, 69]. For example, Whittier et al. deployed an online survey shortly after a popular television show had aired an episode with a story line about syphilis and found that participants who had seen the episode reported higher intention to be screened for syphilis [68]. In contrast, movies can also have harmful learning effects. For example, misrepresented medical scenes can lead to dangerous misconceptions that reinforce racist stereotypes [44] or lead to self-diagnosing with insufficient medical understanding [49]. This has led to the creation of programs for reviewing movie contents [47,60]. Hoffman et al. found that medical television's influence on viewers' health-related knowledge was deemed negative in 11% of prior studies, positive in 32% and mixed in 58% [27].

Unlike the formats that we typically associate with the idea of "learning," people learn *passively* from film media [35]. Krugman and Hartley assert that this type of passive learning is "characterized by an absence of resistance to what is learned" and so in some ways has capacity to be especially powerful [35]. But the precise impact a movie on a particular person is likely unpredictable. Fearing argues that "what the individual 'gets' [from the movie] is determined by his background *and his needs*. He takes from the picture what is usable for him or what will function in his life" [19]. Integrating movies as a tool for active teaching/learning has been widely discussed in fields such as medicine [6, 36, 65], counseling [34], and international politics [18].

Film studies is a rich field whose methods often involve close watching of one or a small set of films, sometimes frame-by-frame analysis, and factoring in how elements such as the film creators' personal backgrounds and societal or cultural issues that may have influenced the film itself and its reception by audiences [13, 56]. The increasing availability of digital analyses, which enables "big data" in film studies, has shifted approaches and spurred, for example, the establishment of the Digital Cinema Studies network [57]. Subtitle analysis has been used to gain insights about the contents of bigger sets of movies. For example, linguistics researchers studied speech acts by analyzing "Evim Sensin" (You Are My Home, 2012) subtitles [29] and compared word frequencies in Greek and Polish movie corpora [15, 40]. These analyses found that language in subtitles is similar to everyday language and

that topics from society are reflected in films. Other research is based on searching for words in subtitles to analyze, for example, hate speech or physical aggression and verbal insults within selected movies [61, 70] or how sex behavior is referenced in a Netflix series [71].

User-Focused Password Research   There is a significant body of Usable Security & Privacy research regarding passwords and technology users, which seeks to answer questions such as: What are users' existing password practices [33, 41, 58, 62]? What do users understand or believe about passwords, password strength, and password attacks [33, 42, 58]? In what ways are passwords typically attacked [50]? How can we encourage users to create better passwords or otherwise decrease vulnerability to authentication attacks [2, 55, 72]? Common practices that make users' accounts more vulnerable include creating predictable passwords [58], re-using passwords across different services [41], and using personal information in a password (e.g., year of birth, names of relatives) [62]. Users also regularly expose these types of personal information online [28]. Analyses of leaked passwords show that users often add numbers at the end of their passwords, capitalize the first letter of the password, and make common letter substitutions (e.g., "@" to replace "a", or "1" to replace "i") [33]. Users tend to overestimate the security of passwords they create [58] and have different misconceptions regrading password composition, handling and attacks [42]. Security researchers have also formed an understanding of how passwords (or authentication systems more generally) are typically compromised. This can happen via automated password guessing (e.g., brute force or dictionary attacks) or compromising other parts of a user's security (e.g., deploying a keylogger or using social engineering to get a user to reveal their password) [50].

3   Method

In previous work so far only a targeted selection of films have been examined and with a very specific focus on hacking, so we take an approach that enables us to draw quantitative conclusions about passwords in movies. To scale our analysis via automation, we used text-based approaches to analyze a large set of movie subtitles.

Creating a Movie Subtitles Dataset   We obtained the subtitles from a torrent link posted on the social news aggregation website Reddit *r/DataHoarder* [3]. The torrent contains a database (`opensubs.db`, 136.8 GB) of 5,719,123 subtitle files, crawled on July 24, 2022. It also contains a metadata file (`subtitles_all.txt.gz`, 309 MB) that includes information such as movie name, year, language, content type (movie, TV show), season, episode, IDs (IMDB, OpenSubtitle), upload date, frame rate, and file format.

The Reddit thread stated that these subtitles were initially sourced from the website `opensubtitles.org` [8], one of the largest subtitle databases on the Internet. Subtitles are uploaded by users, who then vote and comment on the quality of subtitle files.

We filtered out non-English subtitles and subtitles for content besides movies because the full torrent also contained subtitles for TV shows and other content types. Movies that appeared in a non-English language (e.g., Parasite, 2019; Korean) but had English subtitles available were included in the analysis. Additionally, the torrent contained duplicate subtitle entries for some movies (e.g., if two users had uploaded subtitles for the same movie); we removed duplicates by always using the most recently-added subtitle file. Our final dataset contained subtitles and metadata for 97,709 movies. More information about this movie subtitle dataset (e.g., graphs of their genres and years of distribution) can be found online[1]. We obtained additional metadata including genre, popularity, and other details from *The Movie Database* (TMDB), using the TMDB API.

Identifying Password-Related Content in Movies To automatically identify content in movies that is related to our research topic, we perform a search within the subtitles for the word *password*. We found that this straightforward approach was the most appropriate for identifying relevant content in such a large dataset. We considered including other authentication-related words or phrases in our keyword search, including *passphrase* (occurs only seven times in the dataset) and *PIN* (high false positive rate due to semantic overload). *Password* appeared 5,982 times in 2,851 different movies (just under 3% of movies in our dataset).

To create units of analysis corresponding approximately to the notion of a movie "scene," we considered the nine subtitle lines before and after the occurrence of the word *password* (i.e., a total of 19 lines). Note that subtitles do not encode the idea of a "scene;" we found this to be a conservative approximation (i.e., the researchers agreed that 9 lines before/after the keyword were more than enough context to meaningfully interpret the data). Subtitles include newlines corresponding to what would appear as one line of text on someone's screen if they were watching the movie with captions. There is not information about who said which words. A longer dialogue from one character may span multiple lines. Typically (but not always), newlines or other visual indicators such as dashes are inserted when a new character begins speaking. Subtitles typically contain (most of) the spoken words, though crosstalk (i.e., multiple people speaking at once) and background dialogue may not be fully captured. Sometimes subtitles contain additional information about the audio such as "laughter" or "music". In the results, we report on patterns of how these

movies with password-related content are distributed in terms of year and genre.

Characterizing Scenes about Passwords We started with an open-coding approach to analyzing these 5,982 password-related scenes. Two authors each used MaxQDA to independently open code the same set of 50 scenes, which included 10 randomly selected scenes from each of five time intervals (including very old and very recent movies). The researchers then compared their open codes and generated a codebook. One author applied the codebook to all scenes. When coding decisions were unclear, he consulted with co-authors to reach a consensus. The codebook can be found online[1] alongside with a list of all scenes including the movie metadata and the set of codes we applied to each scene.

Analyzing Password Topics and Password Attacks We characterized the context of use for the password (e.g., if the password is used for a computer, a website account or locks), and different activities that are performed with passwords (e.g., password creation, change or losing a password). In particular, we coded the scenes based on whether they contain *password hacking* and/or *password guessing*. These codes were used as the basis for generating a sample of movies that we watched manually. We report summary statistics and patterns that emerge between these codes and also over time/by genre, and we include relevant quotes from the subtitles to illustrate our findings and provide qualitative depth.

Measuring Movie Passwords' Strength While applying the codebook, we recorded all passwords that showed up in the transcripts (e.g., *"Your password's 999999?"* (Max Winslow and The House of Secrets, 2020)). In cases where passwords were described verbally, we recorded our best approximation of the plaintext passwords (e.g. in (The Disappearance of Jennifer Dulos, 2021), subtitles state *"What's the password? - It's four zeros"*; we recorded this as `0000`). This resulted in a list of 687 passwords which are listed in Appendix A organized by strength using the *zxcvbn* metric as described below. To measure the strength of these passwords, we applied two well-known password metrics:

- *zxcvbn*. A simple but relatively accurate [22] strength meter developed by Dropbox [66]. *zxcvbn* categorizes passwords into 5 strength categories from weakest (Class 0) to strongest (Class 4). Passwords up to Class 2 are easily guessable, whereas Class 4 contains passwords that would require 10^{10} guesses [33].
- *PGS*. The "Password Guessability Service" created by researchers at Carnegie Mellon University [59]. The output of this metric is the number of guesses it would take to guess the password (or -5 if it cannot be guessed).

We compared the strength of these passwords with real-world leaked passwords from several well-known breaches:

- The Popular-200 (200 most used passwords of 2023) [43], a current dataset with a focus on frequently used (and therefore tending to be less secure) passwords.
- The *Ignis 1M* wordlist [25], which was assembled in 2020 from various data leaks (Collection #1, Dropbox, LinkedIn, and others) and contains the 1 million most popular passwords found within those data leaks [32].
- A list from the *RockYou* data breach [14], which was leaked in 2009, but contains the full distribution of passwords from the weakest to the strongest since it was leaked in plaintext and is frequently used in comparable research.

Watching Movies to Gain Deeper Qualitative Insights   To find additional information that could not be found out by analyzing only a specific scene (such as the importance of the password activity for the whole movie) or could not get out of the subtitles (such as a password which is typed in but never said out loud, which is why it may not appear in the subtitles) and to compare the findings of our subtitle analysis with some real movie scenes, we watched a small subset of 21 movies and evaluated the password attack scene in the context of the overall movie plot. Considering movies that came out between 2013 and 2022 which were sorted from most to least popular based on the total amount of votes a movie reached on TMDB [5], we included the top 10 movies that contain *password hacking* and the top 15 movies that contain *password guessing*. Three movies were in both categories (i.e., contained hacking *and* guessing scenes), and one movie was not available to watch online.

Six people participated in this task, watched the allocated movies in full, and filled out a questionnaire that was discussed and iterated on by the authors; the questionnaire aimed to capture details that would have been missed in transcript analysis. All six people who participated in this task were trained and had opportunities to ask questions about their understanding of the questionnaire before starting. The questionnaire and list of watched movies can be found online[1].

High-Grossing Movies   Our dataset includes both very popular and relatively obscure movies. Popular movies have (by definition) already reached a broader audience and will likely continue to be viewed more often, which makes their capacity to (mis)inform viewers especially important to consider. While our analysis is primarily concerned with the full dataset, we assembled a secondary High-Grossing dataset to assess whether high-grossing movies' characteristics are meaningfully different. The High-Grossing dataset consists of all movies in our dataset that appear in the list *Top 1000 Highest-Grossing Movies of all Time* [7] as of April 2024. This list contains 39 movies that are newer than our dataset and three that are not available in English (i.e., excluded from our dataset). Of the remaining 958 movies, 70 (7%) contain

the word *password* at least once (listed in Appendix C). We compare the High-Grossing movies with our other findings in Section 7.

Ethical Considerations   Movies are subject to copyright; however, analysis like ours should be protected under fair use. Additionally, the user-generated subtitles on OpenSubtitles.org seem to not infringe on copyright [64]. While OpenSubtitles.org disallows scraping, they explicitly allowed non-registered users to download subtitles at the time of database creation and still allows non-commercial, scientific, and educational use [64]. An older corpora from this site was published in 2018 and is used by the linguistics community [37]. By using a dataset that had already been scraped and shared publicly, we avoided stressing OpenSubtitles.org's server bandwidth. Finally, while the subtitles are user-generated, we did not use data about any of the individuals who uploaded them.

Limitations   Some of our methodological choices present limitations to how readers should interpret our results and what we could find. However, these trade-offs enabled for a much broader analysis than has been conducted previously, and so represent a deliberate choice. Searching for only the keyword *password* almost certainly excluded relevant scenes and movies (both content that is relevant to passwords specifically, but also content more broadly related to authentication or security and privacy in general). Using only subtitles limits what information we can analyze, and we must assume that the subtitles sometimes leave out relevant context or are misleading. Our deep qualitative analysis through watching a small number of movies helps address this concern (see Section 5). It is possible that we could have understood the scenes marginally better by including more than 19 subtitle lines in our analysis, though we found this to be acceptable empirically. The subtitles used in this work were user-contributed (on opensubtitles.org [8]) and some of the subtitles were translated, which may change the meaning of individual sentences. We only informally checked the data quality (by paying attention to subtitles during the movie watching activity), but we found that the subtitles were highly accurate. Finally, while this paper is the largest study on how cybersecurity-related topics are presented in entertainment media, we do not consider other types of content besides movies, including TV shows (or series) or other online media.

Additionally, it must be emphasized that our analysis focuses exclusively on the depiction of password topics in films and thus on the question of what people are shown when they watch the movies - and could potentially learn from them. Whether and to what extent people actually do take away this information from films remains to be investigated. Therefore, there is no evidence presented that these scenes have any real-world impact.

4 Results: Depiction of Passwords in Movies

Next, we present the results of our analysis. In the first step, we will look at which movies contain password-related content at all, before turning to different scenes and what is done with passwords in them. This includes which things passwords are used for, which activities are described in the context of the passwords, and more.

4.1 Movies Featuring Passwords

Recall that only around 3% of movies in our subtitle data set contain the word *password* (2,851 movies). We start by asking about the characteristics of these movies. Are passwords more likely to be mentioned in certain genres of movies? How has the frequency of mentioning passwords changed over time?

Figure 1 shows how movies whose transcripts include *password* are distributed across various genres are more likely to contain *password*. *Password* is mentioned most commonly in Thriller, Science Fiction, and Action movies and least frequently in Western, Music, and Documentary movies.
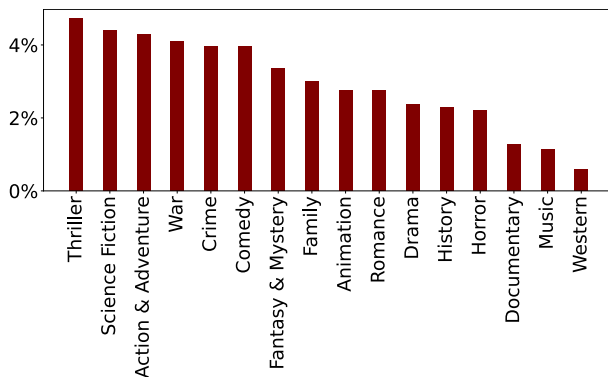
Figure 1: Percent of movies within each genre that contain the word *password* at least once. Thriller is the genre that mentions passwords most often, with 5%.

As shown in Figure 2, the proportion of movies containing *password* increased over time which means that newer movies are more likely to mention *password* than older ones. For movies that have come out since the start of 2020, slightly over 5% (or 1 in 20 movies) contain the word *password*. 71% of movies with *password* have been released since 2005.

Comparing the High-Grossing movies with the overall dataset, a comparatively high number contain password scenes (7%, compared to 3% in the overall dataset or 5% of all movies since 2020). This may be explained by the makeup of the High-Grossing dataset: they tend to be newer (545 came out after 2010) and they tend to fall into genres that we found more often reference passwords (78% are Thriller, Sci-Fi, or Action & Adventure). We return to this comparison in 7.
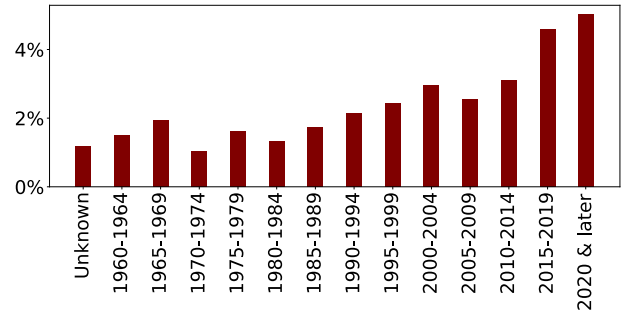
Figure 2: Indication of what percent of movies within a year interval contains the word *password* at least once. In recent years, slightly more than 5% of films contain the word *password*, in past years this was between 1% and 2%.

4.2 Password Behavior in Movie Scenes

Next, we present the types of password behavior that we detected within the analysed movie scenes as well as patterns that emerge based on applying these codes to the data.

Context of Use   We were often able to use the subtitles to discern what the context of use was for the password referenced in a particular scene. We categorize these contexts into computer-related, Internet-related, banking, (interpersonal) legitimation, or anything else, as shown in Figure 3. A more detailed breakdown of contexts of use is contained online[1]; here we primarily report findings related to these high-level categories. *Interpersonal legitimation* contains all scenes in which a person uses passwords or passphrases to prove towards other people that they belong to a specific group or are allowed to perform a certain activity (e.g., enter a restricted area, perform an operation as a spy, etc.). Example scenes are the following: *"It had to be a Gryffindor. Nobody else knows our password"*(Harry Potter and the Chamber of Secrets, 2002); *"you'll get your instructions day by day. The password for the contact will be »Wee-wee Birdie«, and the contact will answer »Poo-poo birdie«*(Brigada explosiva: Misión pirata, 2008). The most common *Internet-related* passwords are Wi-Fi passwords (including passwords *"to the internet"* (Witness to Murder, 2019)) and passwords for website accounts such as the *"registration in a site for dating"* (Love.net, 2011), or for the *"website you have an administrator's account, right?"* (Suicide Club, 2018). *Other* combines a collection of all scenes that do not fit into the other categories. This includes, for example, when a password is used as a signal to start a specific activity as in *"At the password, »The cat is in the kitchen cupboard«, you'll open the envelope"* (In Danger and Dire Distress the Middle of the Road Leads to Death, 1974). Other examples in this category are when a password is used to control the people's minds or (in one case) pets, or when it is used as a mantra such as *"and remember the password: relaxation"* (The Big Bluff, 1995).

We observe shifts in the most frequent context of use over time; until 1990-1994, the dominant context of use was legitimation. In more recent time intervals, digital passwords used in computer contexts have become most common, and from the 2015s onward, Internet-related passwords have been dominant.

	Unknown	1960-1964	1965-1969	1970-1974	1975-1979	1980-1984	1985-1989	1990-1994	1995-1999	2000-2004	2005-2009	2010-2014	2015-2019	2020 & later
Computer	9	0	0	2	1	17	15	24	80	110	129	203	396	171
Internet	8	0	1	0	0	0	2	4	21	27	49	142	402	222
Banking	0	1	0	4	0	1	4	4	13	39	31	59	122	36
Locks	0	0	0	0	6	0	3	4	5	17	14	16	31	16
Legitimation	111	26	57	33	50	39	48	40	39	65	103	89	136	54
Other	19	3	10	5	7	6	8	13	38	44	64	66	187	70

Figure 3: This figure shows how the context of use varies over time. Darker cells indicate a larger number of scenes with this context of use and release date, and the number in each cell conveys the number of scenes. Notice that legitimation was more common than other contexts in older movies, but computer- and Internet-related contexts have become most common recently.

Password Life Cycle In some scenes, characters speak about specific points in a password life cycle. Our coding process distinguished: password creation, changing a password, training to remember a password, losing or forgetting a password, performing a password recovery or reset, and reusing a password for different accounts. The distribution of these in scenes over time is shown in Figure 4. More detailed descriptions and examples of each point in the life cycle are included in Appendix B. Except for password recovery and password reuse, all codes appear with similar frequency in the movies; they are named between 104 and 115 times.

	Unknown	1960-1964	1965-1969	1970-1974	1975-1979	1980-1984	1985-1989	1990-1994	1995-1999	2000-2004	2005-2009	2010-2014	2015-2019	2020 & later
Creation	7	2	1	0	4	3	5	2	5	4	17	15	29	10
Change	1	0	2	0	1	0	2	2	2	3	11	19	47	25
Training	9	2	4	4	5	1	4	4	3	8	9	12	33	10
Forgetting	4	0	5	1	2	3	0	3	6	8	7	11	35	29
Reset	2	0	0	0	0	0	0	0	5	1	0	5	19	10
Reuse	0	0	0	0	0	0	0	0	0	1	0	4	9	7
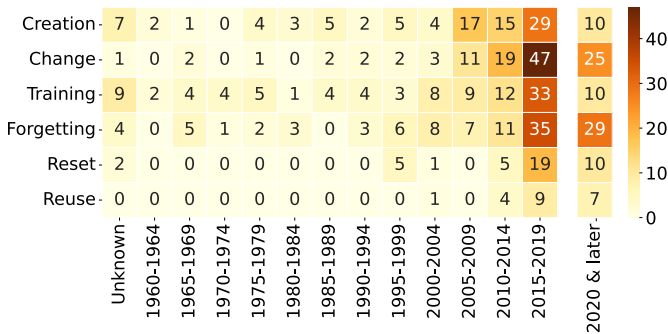
Figure 4: Points in the password life cycle and the number of scenes they occur in over time. Darker cells correspond to a larger number of scenes.

Password Sharing A password is shared in scenes from more than 1000 movies. Here, we consider *how* this happens and *with whom* the password is shared.

Intended/intentional sharing is the most common type of password sharing, presented in 469 scenes. Of those, 74 scenes include a deeper explanation of how the password should be handled, such as, *"Once you pass through the first step, you will receive a password on your phone. The last step is the key."* (Collectors, 2020). However, password sharing is often non-consensual or forced (132 scenes). For example, *"The silver bowl your brother-in-law got from Turkey... Do you know how much its worth? Do us a favor, Just give the password to the lock of all the precious things in the house"* (French Biriyani, 2020). In 64 scenes, a password is shared unintentionally, such as in the following scene: *"You shouldn't leave shit lying about. -How'd you get the password? -You had it taped underneath the fucking thing."* (Boy A, 2007).

In 367 scenes, the password is distributed to one legitimated person such as a friend, colleague or family member. In most of these cases, the recipient takes on the role of a friend (186 scenes), followed by family members (78 scenes), work and the partner. In 93 scenes, it is shared with a small group of people, e.g., some direct colleagues or people from the same squad. In nine scenes, a large group is the recipient. For example, a password is forwarded via radio to all military units or *"the whole FBI"*(Enemies of the State, 2020).

Security Best Practices As described in the background section there is a large amount of security advice regarding password behavior. In addition to general recommendations such as not sharing passwords (discussed above), using password managers and multi-factor authentication are recommended as specific best practices. We have therefore analyzed what role they play in movies.

Password managers appear in only four scenes, and all four scenes are in one movie. It is the French movie *Disappear: Cover your online tracks* from 2021, a documentary including (among several other topics) a description of what a password manager is and how it can be used.

In seven movies – all from 2013 or later – the password is combined with a one-time password (OTP). That is, Multi-Factor Authentication is shown in these movies. In four scenes, the OTP is sent to a smartphone, in one to a key card, and in one it is created by *"a pair of watches that had undergone a special magnetization process. Only by putting the two watches together will a person be able to acquire the correct account and one-time password, thus, gaining access to the huge sum"* (Arjun Suravaram, 2019). In four scenes, the OTP is used to access a bank account or transfer money; in one, it is used for accessing a high-security area in a building, and in another, it is used to perform a password reset via phone (the OTP is sent to the phone and has to be read aloud).

5 Depiction of Password Attacks

In the coding process, we distinguished two basic kinds of password attacks: *Password guessing*, where a human actively guesses candidate password based on frequent passwords, specific knowledge about a person, or known old passwords, and *password hacking*, where other techniques are used to obtain the password such as social engineering or shoulder surfing or using automated tools for (brute-force) guessing. Password guessing appears in 220 scenes and password hacking in 63.

Password Guessing We observed different guessing approaches: Sometimes people try *typical passwords*, hoping that the target chose one of the easily guessable ones, e.g. *"tell him that 1-2-3-4 as a password is worth fuck-all"* (Todos tus secretos, 2014). Others use their knowledge about the person using the password and try to guess it by characteristics of the target, e.g. *"The name of her boyfriend is Troy. But she calls him Batchoy. Her birthday is July 6th. So let's try this"* (Walwal, 2018). Sometimes, attackers know old passwords and hope they have not been changed: *"Paul's used the same combinations and passwords since we were freshmen at Cornell"* (Consensus Reality, 2018). In some cases, the attacker looks around the physical surroundings and searches for (physical) clues or written-down passwords, for example *"she scoured their apartment looking for passwords to get into his laptop"* (Trust No One: The Hunt for the Crypto King, 2022). In other scenes, the attacker just tries whatever passwords come to mind. Most frequently, people in the movies use *trial & error* (94 scenes, 43% of all scenes with password guessing), closely followed by using knowledge about the person (83 scenes, 38%).

Overall, the guessing is successful in 103 scenes (47% of scenes with password guessing), 85 attempts fail (39%) and 32 (15%) have an uncertain outcome. Figure 5(a) shows guessing attempt success broken down by approach. Attacks based on knowing an unchanged password have the best success rate (11 out of 13, 92%). Using "typical" passwords and targeted guesses had a higher-than-average success rate (76% and 53%). Trial & error and using the physical surroundings have higher than average *failure* rates (54% and 58%, respectively).

Password Hacking We define *password hacking* to include all attack techniques except *guessing* (e.g., social engineering, shoulder surfing, overhearing passwords, using malware, virus software, keyloggers, or password cracking tools). In applying our codebook to the subtitles, we categorized the 63 scenes with hacking attacks into: (1) 19 scenes that focus on the human element (i.e., overhearing, shoulder surfing, and social engineering), which occurs in around 30% of these scenes; (2) 31 scenes that focus on the computer (i.e., breaking passwords, using master passwords to circumvent individual passwords, virus, brute-force, etc.), in almost 50%

	Typical PW	Targeted Guess	Unchanged PW	Trial & Error	Physical Surrounding
Success	13	44	11	32	3
Failure	4	22	1	51	7
Uncertain	1	17	1	11	2

(a) Guessing Attacks

	HUMAN ELEMENT			COMPUTER FOCUSED						OTHER
	Overhearing	Shoulder-surfing	Social Engineering	Crack/Break PW	Master PW	Technobabble	Steal PW	Circumvent PW	Brute Force	Other
Success	6	7	4	4	6	4	2	7	3	8
Failure	0	0	1	1	0	2	1	0	0	4
Uncertain	1	0	0	0	0	0	0	0	1	1
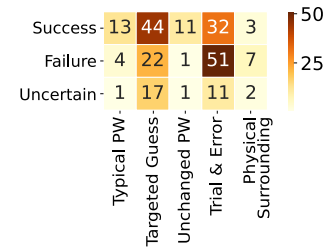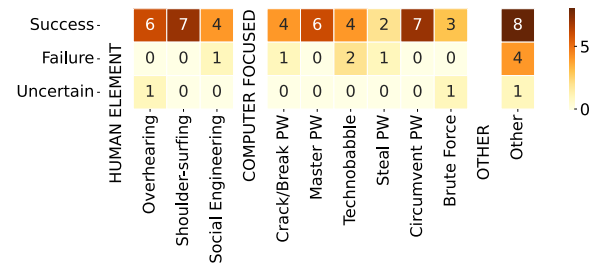
(b) Hacking Attacks

Figure 5: Showcasing the amount of cases where codes applied to (a) guessing or (b) hacking and authentication outcomes overlap thus highlighting which attack strategies lead to which outcome of authentication. Cells are intensifying in shades of red proportionally based on the prevalent guessing strategies authentication outcome.

of hacking scenes, and (3) other in 21% of scenes. Hackers often performed some kind of "computer magic," described with *technobabble* in the movie dialogue, as in the following scene from (America: The Motion Picture, 2021): *"After a reverse hash, I backdoored the root password. A base checksum against the main data store allowed me to retrieve the salted hash, and then, from there, I was gleaming the cube."*

Password hacking in movie scenes tends to be even more successful than password guessing (81% success, 14% failure, 5% uncertain; see Figure 5(b)). Human-factor attacks are successful in 89% of scenes (17 of 19 successful attacks).

Deeper Insights from Watching 21 Movies with Hacking and Guessing Here we convey our findings from watching 21 full movies that contain password hacking and/or guessing. We examine three key details that we were unable to analyze based on subtitles alone: the roles and character traits of attackers and their targets, more nuanced understandings of the assets targeted in password attacks, and the importance of the attack to the overall plot of the movie. We also consider the extent to which our subtitle analysis may have been incorrect or incomplete within the factors included in our codebook.

In movies, there is generally a clear distinction between "good guys" and "bad guys," main and secondary characters, and it is generally easy to understand aspects of character development such as character traits (e.g., computer exper-

tise) and relationships between characters. While it was too difficult to discern these roles and character traits based on subtitles alone, we now paid special attention to these. Out of the 21 movies we watched, the character trying to hack or guess a password was only "bad" in two movies (i.e., in 90%, the person doing the hacking was good or neutral). In movies with hacking, the character(s) performing the attack were always main characters (if hacking occurred in a team, at least one team member was a main character); the characters trying to guess passwords were mostly of average importance (e.g., a friend or family member of the main character). Characters (or at least one team member) performing hacking had high computer knowledge, but those trying to guess passwords mostly did not have high computer expertise. These characters included: two superheroes, one police investigator (assisted by Batman), and a couple of gangsters, one of whom is a hacking specialist. The targets of password hacking attempts were typically opponents (good or bad) of the attacker, with no close relation except being rivals. Password-guessing targets were only active opponents in two cases, and they included family members (5 movies), characters with a romantic background (2 movies), colleagues or friends (2 movies), and neighbors (1 movie).

By watching movies, we were also able to better understand assets targeted by guessing or hacking attacks. Subtitles often mentioned only *the computer*, but we could not assess what role *this computer* plays in the plot. Within the movies we watched, we observed that hacking tended to target civic or company assets. For example, civic assets include targets in the context of secret agents and similar, which are of great (civic) importance such as the computer system of S.H.I.E.L.D.,[2] secret online videos concealed by criminals or a city's traffic control system. Company assets that were targeted include the code to the space station, company servers with precious software, or employee data. In contrast, password guessing tended to target private assets, and the targeted assets themselves were of little importance to the plot (e.g., private computers and smartphones, two email accounts and once "all my private accounts"). Password hacking tended to be of high importance for the story (in all but one movie with hacking). For example, successful hacking saved thousands of lives in *Captain America: The Winter Soldier* (2014), circumvented the next crime in *The Batman* (2022), and resolved the entire plot of *Focus* (2015), a movie in which getting the secret code constitutes the main story line. Password guessing was mostly less important to the plot. For example, in *Blended* (2014) the main character's son finds out she was on a date, but he would have found out a bit later anyway. We rated the password guessing as having "medium" importance for only two movies; in both cases, the guessing is only one of several steps to reach the final goal (e.g., freeing the second

main character, who then helps to finalize the next quest in *Ready Player One* (2018)) .

Finally, we specifically sought a deeper understanding of whether our analysis of subtitles alone was misleading or incomplete within the set of topics included in our codebook. Overall, we found little evidence that our subtitle analysis was insufficient. The six specific passwords that appeared in movies we watched were also present in the subtitles, though in some cases, the password appeared on screen before it appeared in the subtitles. For password guessing attempts, no tools or computer activity was ever shown; instead, in most cases characters inform about the attack only afterwards, which is entirely available in the subtitles. The flashy tools and techniques used in password-hacking scenes were more impressive visually, and sometimes the amount of time spent showing this on screen seemed disproportionate to the fraction of subtitles spent describing it. For example, we watched characters bypass the password authentication by using another authentication method (retina scan), brain-to-brain transfer, artificial intelligence as a hacking tool, and fancy illuminated computer screens and tools (without any understandable computer activity shown). However, we found the key points were also understandable without video.

6 Password Strength in Movies

As described in Section 3, we systematically recorded any passwords that were directly stated in scenes (n = 689). We analyzed and then evaluated their strength according to *zxcvbn* (Figure 6) and *PGS* (Figure 7). In this section, the results regarding password strength in the movie database will be shown as well as their comparison to real-world passwords.

Per the *zxcvbn* strength metric, well over 70% of movie passwords are rated as Class 0 through Class 2, meaning that they are easily recoverable (i.e., weak). Within these classes, only the 200 Most Popular contains a higher percentage of Class 0 passwords, though this is expected since it by definition excludes uncommon passwords. On the other hand, the two lists that contain real-world passwords, RockYou and Ignis-1M have a much smaller percentage of passwords in Class 0. RockYou has almost twice as many passwords as movie passwords in Class 2 and the same holds for Class 3 passwords. Finally, it is interesting that the two lists that perform best for Class 4 passwords are RockYou and movie passwords. Since RockYou is a complete dataset (the service stored all of its passwords in plaintext) it represents the most accurate distribution of the strength spectrum of real-world passwords and Figure 6 suggests that the percentage of movie passwords classified as Class 4 resembles a real-world distribution very accurately.

Using *PGS*, we see that most passwords fall within the range of 10^3 and 10^8. The top five strongest passwords from our data set include: `Cv'qrPo` (Our Happy Holiday, 2018), which can not be cracked; `ldfvarumellamsheriaavum`

---

[2]"Supreme Headquarters International Espionage Law-enforcement Division," a fictional counter-terrorism intelligence agency from the Marvel cinematic universe (we watched Captain America: The Winter Soldier, 2014).

(Varane Avashyamund, 2020); `T19FXP07YT567TZ5` (Those Who Are Fine, 2018); `DOOMEDIFYQUQUIT` (Sono tornato, 2018); and `Youwereneverthereformed@d` (FML, 2016).
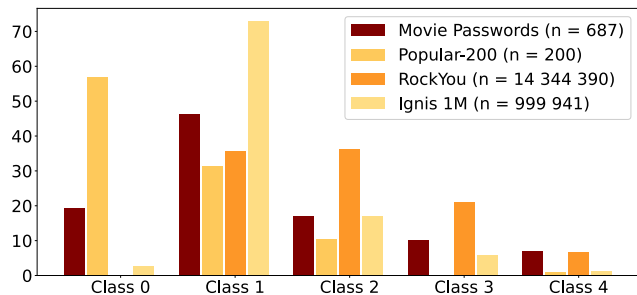
Figure 6: Comparison of the strength distribution of passwords found in movies (movie passwords) and control datasets (Popular-200, RockYou, and Ignis 1M) using the *zxcvbn* classification. The weakest passwords are in Class 0, the strongest in Class 4. The movie passwords contain a distribution closely resembling RockYou with a significant number of passwords belonging in Class 4.
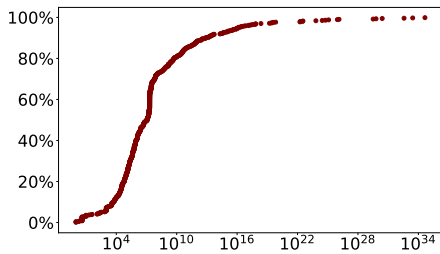
Figure 7: The guessability of passwords found in movies using *PGS*. On the x-axis the number of guesses (log scale) is charted, on the y-axis the percent of how many passwords from within the dataset are guessed.

Are Passwords in Some Contexts Stronger?   Returning to our previous context-of-use analysis, we find that the distribution of password strengths is not uniform across all contexts of use (Figure 8). Computer-related and Legitimation passwords have disproportionately more Class 0 passwords, and Internet-related passwords are the context of use with the highest percentage of very strong (Class 4) passwords. Locks, have disproportionately weak passwords – 96% are in Class 0 through 2, which indicates they are easily recoverable, and none are in the strongest Class 4.

There are 73 scenes for which we were able to determine a life cycle point *and* in which a password was directly stated (i.e., for which we can measure the strength of the password in that scene). The majority of passwords in each context had low security (Class 0 or 1, per *zxcvbn*); in password

recovery and reset scenes, 71% were in one of these two lowest-strength classes. The strongest passwords occurred in the context of losing and forgetting a password, where 30% of passwords were Class 3 or 4.
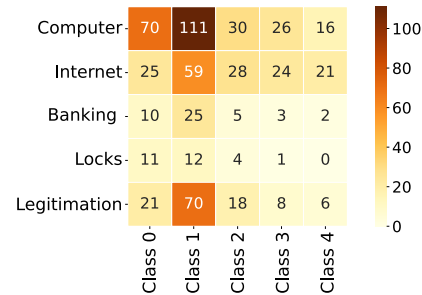
Figure 8: Showcasing the number of cases where password strength according to *zxcvbn* classes overlaps with the different contexts of use. The cell color indicates the relative relationship to other classes by use case with shades of yellow intensifying relative to the code usage in the interval. Strong passwords (Class 3 and 4) are represented more frequently for computer and internet-related topics.

We take a closer look at the more detailed breakdown of Internet-related contexts of use (Figure 9), since they have an especially high percentage of strong passwords. Email and Wi-Fi passwords have an atypical distribution: Passwords of Class 4 are more common than of Class 2 and 3. Streaming/Cable Account passwords stand out as well: The two passwords found for this specific use case are of Class 3 and 4 which makes this category the "strongest" of all categories investigated. However, the small number of passwords found must be taken into account.

This results indicate, that there are certain topics, where strong passwords are considered typical, including email, Wi-Fi, and streaming or cable accounts. In other areas such as locks weak passwords are almost always used.

Are Weaker Passwords More Susceptible to Attack? 133 of the scenes with password guessing (60%) and 10 of the scenes with password hacking (16%) include a specific password whose strength we can analyze. 61 (46%) of password guessing attempts were successful (40% fail and 14% unclear). On the other hand, 100% of the hacking attempts were successful. Both of these success rates are relatively similar to the overall success rates for password guessing and hacking (47% and 81% success, respectively).

Observing how guess success rates differ within the five *zxcvbn* strength classes, as shown in Figure 10, it can be observed that success or failure are only closely related to the strength class. Generally speaking, among low and high classes success or failure of the guessing attack are almost equally likely. Exception is class three where 60% of the attacks are successful.

Multiple Passwords - 0 6 0 0 0
Website Account - 6 9 7 4 2
Email Account - 5 6 3 3 4
Wi-Fi / Internet Connection - 9 22 12 12 13
Social Media - 0 11 6 3 1
Streaming / Cable-Account - 0 0 0 1 1
Telecommunications / Pager / Voicemail - 5 5 0 0 0
Cloud - 0 0 0 1 0

Class 0   Class 1   Class 2   Class 3   Class 4
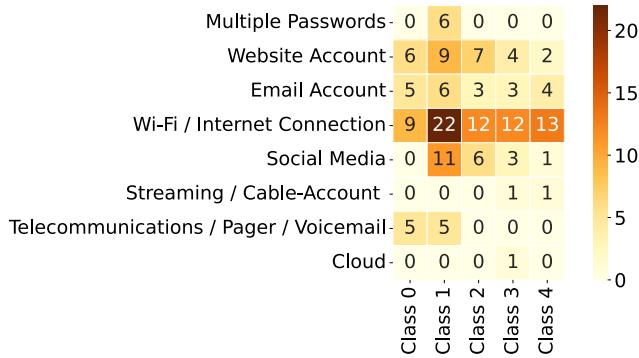
Figure 9: Showcasing the number of cases where password strength according to zxcvbn classes overlaps with the different internet-related contexts of use. The cell color indicates the relative relationship to other classes by use case with shades of yellow intensifying relative to the code usage in the class. Certain usages show a polar distribution of either very strong or very weak passwords.

Success - 50% 41% 33% 60% 50% 46%
Failure - 43% 39% 33% 33% 50% 40%
Uncertain - 7% 20% 33% 7% 0% 14%

Class 0 (133)   Class 1 (317)   Class 2 (118)   Class 3 (70)   Class 4 (49)   Total
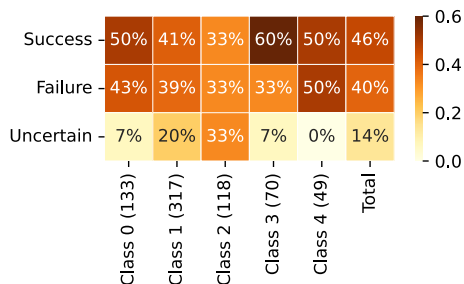
Figure 10: Relation of success, failure and uncertain outcomes of password guessing attacks among different password strength classes according to *zxcvbn*. The cell color indicates the relative relationship between outcomes of guessing attacks and password class with shades of yellow. It is outstanding that for Class 3, the success of the attack is more likely than its failure.

7   Discussion and Future Work

We found that passwords are increasingly being mentioned in movies (since the start of 2020, around 1 in 20 movies mentioned passwords); thus, it is especially important and timely to consider how realistic they are, what this might be teaching people, and what should be done to create a framework for the dissemination of security topics.

Are Passwords Portrayed Realistically in Movies?   Our analysis involved directly comparing the strength of passwords from movies to those leaked in real-world data breaches (see Section 6). Movies contained passwords with a wide range of strengths, just like those in the real-world data sets we compared against. While the patterns we observe depend on exactly which data set we compare to, we found that,

broadly speaking, movies contain more of both especially weak *and* especially strong passwords than most data sets. Interestingly, the fraction of Class 4 passwords (i.e., strongest according to *zxcvbn*) in movie passwords most closely resembles those in the RockYou data set, which is the only full leaked list of passwords we could compare with, so is in some ways the most realistic.

We can also postulate about the realism of other findings. For example, movies do seem to portray realistic contexts of use (e.g., ranging from car locks or computers to email and streaming accounts), real points in the password life cycle (e.g., creation, change, forgetting, and resetting), and realistic behaviors such as password sharing and reuse. Additionally, the trends in common contexts of use over time seem to have approximately shifted with the evolution of technology: computer-related uses of passwords picked up in the early early 1980s, and Internet-related contexts started catching up or taking over in the mid to late 2010s.

The realism of password hacking and guessing in movies is somewhat of a mixed bag (see Section 5). Out of the total 283 scenes with either type of attack, over half of them were successful which confirms the literature which states that attackers are too powerful [23, 38]. But the *way* password guessing is portrayed *is* quite realistic: characters use approaches such as trying out typical passwords, using birthdays, hobbies and pet names, or hoping that a person has not changed an old known-to-the-attacker password. Prior work has shown that these types of knowledge about a person can help with guessing their passwords [31]. Guessing often happens in the context of family and friends, which (correctly) shows that attackers are "not only criminal hackers but also people you know" [12]. The small number of attempts needed to guess passwords seems unrealistic, but this may be related to directors' desire to keep movies from becoming dull.

Unlike prior studies [23], we found that hacking scenes do often have some basis in reality. For example, we found instances of real-world tools such as keyloggers being used in the attack, and we found that attacks exploiting a human factor were common (around 30% of attacks) and more successful than other attack types. Many of the assets targeted in these scenes were also plausible (e.g., company servers as in *Focus* (2015) or illegal platforms investigated by the police as in *The Batman* (2022)). Still, many of the hacking tools and mechanisms shown were unrealistic (e.g., connecting brains, using a hacking artificial intelligence, etc.) and important situations are missing, such as attacking people who are not rich or in particular significant positions, which may lead to the feeling that one is not "important enough to be targeted" [12].

Are Passwords Portrayed Differently in High-Grossing Movies?   As stated in Section 4, High-Grossing movies were more likely to contain the word *password*, which is perhaps related to the fact that they are more likely to be newer movies and more likely to be in the genres Thriller, Sci-

Fi and Action & Adventure. Appendix C contains statistics and figures comparing the High-Grossing dataset with the overall dataset. Though we have not performed statistical comparisons, the major patterns are largely consistent with the overall dataset, but there are some interesting differences. There is much more likely to be uncertainty about the success or failure of password guessing or password hacking in High-Grossing movies (around 50%, compared to less than 15% in the overall dataset). Of those scenes where the outcome is known, there is a higher chance in High-Grossing movies that the attack was successful; excluding uncertain outcomes, 80% of hacking and guessing attempts in High-Grossing movies are successful compared to only 62% in the full dataset. High-grossing movies are much less likely to depict the Change or Training phases of the password Life Cycle (18% compared to 44% in the overall dataset), much more likely to show a password Reset (24% compared to 8% in the overall dataset), and somewhat more likely to show password Creation (29% compared to 21% in the overall dataset). None of the seven movies that contained multi-factor authentication were in this High-Grossing dataset, nor was the one movie that showed a password manager.

What Are People Likely to Be Learning from Movies?
While our study scope is focused on the contents of movies, we know from prior work that movies can influence viewers' understanding of cybersecurity topics [4, 20, 73]. Our findings suggest that there are both good and bad security and privacy practices in movies that viewers may be learning from, and we find that some security best practices are rarely shown (i.e., it is implausible that viewers would learn to follow these based on watching movies).

We focus here on two key positive practices that viewers could take away from the movies in our data set. First, even though many of the passwords included in movies are somewhat weak, contain personal details such as birth dates that are known to make passwords more guessable [31] and/or are easily guessed by characters, we have hope that many of these scenes may actually be teaching viewers about the characteristics of weak passwords. Most simply, when a weak password is shown as being easily guessed, perhaps this is a cue to viewers that the password is weak. We also observed that characters often make fun of bad passwords in movies, which provides even more direct commentary to viewers. Second, there *are also* many strong passwords in movies. We found that the strongest passwords were used in Internet-related contexts. Within this relatively broad category, strong passwords were especially commonly used for email, Wi-Fi, and streaming or cable accounts. We expect that these scenes could help normalize the use of strong passwords.

As expected, many aspects of our findings point to troublesome lessons viewers might glean from movies. Password sharing and reuse are portrayed as normal behaviors, even super villains use `12345` as their password, and even the boss

of an IT company puts his password under the keyboard. The inclusion of some of these in movies may be justifiable – for example, password sharing involves interaction between people, which plays well to getting multiple characters involved in a scene, but movies rarely spend much time following one character all alone and doing things that might involve passwords (which we must admit, are quite boring). However, even if these behaviors *are* realistic, normal and somewhat justifiable, it is likely still harmful for viewers to see them normalized in movies. Some of the weakest passwords were used for locks or banking, which are high-risk contexts (in fairness, many of these were number PINs, which are realistic and more guessable due to their short length).

Returning to the overall unrealistically high success rate of hacking or guessing passwords in movies, which is even higher in High-Grossing movies, we found that the strength of a password has practically no effect on security. Strong passwords are guessed just as often as weak ones (or even more often) and even for the most secure passwords of Class 4 that are very difficult to guess in the real world, passwords such as `Stephanie'sdude2016` are guessed correctly within seconds. Combined, these portrayals might send the message to viewers that attacks will be successful regardless of security efforts, so why bother trying? In High-Grossing movies, we observed that the outcome of a hacking or guessing attempt was more often uncertain compared to the baseline. We hypothesize that this could influence viewers to see cybersecurity as unapproachable and mysterious, further contributing to tendencies to avoid learning about it and taking appropriate security measures in their real lives.

Finally, we found fewer than 0.3% of movies that mention the use of password managers and/or multi-factor authentication, which are widespread and commonly suggested as part of security best practices [2]. This presents a missed opportunity for movies to help familiarize viewers with these tools.

Implications for Film and Policy Makers, Educators, and Researchers   Because our findings show that many movies portray passwords in ways that could lead viewers to riskier security and privacy behaviors, this paper underscores the importance of recommendations from prior work that call for the creation of a "Cybersecurity in Entertainment Task Force" to consult with both security experts and film makers to help ensure that portrayals of passwords (or of technology more broadly) does not lead to harmful negative outcomes for viewers [20]. Such consulting efforts have already been successful in other domains, such as medicine. Additionally, we contribute an understanding of what contexts of use, password-related behaviors, plot dynamics, and misleading or problematic portrayals have been most common in movies so far. This could help consultants tailor what topics they are most prepared to consult on, and it could help them guide film makers to decisions that are less cliche.

Prior work has also emphasized the capacity for educators to leverage movie scenes in their lessons [20]. We agree and suggest that these could help engage students and enhance their understanding of the content. Studies have shown that the inclusion of movie clips in other educational is promising [18, 34, 39], but should be approached with care when the clips contain inaccuracies [10]. As stated in Section 3, we released a database of our findings (i.e., codes) for each movie in our data set. This can be a helpful resource for cybersecurity educators to find the most relevant clips. Additionally, our findings can help guide curriculum development and instructor decisions regarding which topics are most important to cover (i.e., perhaps focusing on topics that our work suggests students are especially likely to be misinformed about).

Finally, our findings motivate future work in this research direction. As discussed in our limitations section, it is likely worthwhile to study other forms of entertainment media in similar ways. For example, television shows (or series) follow the same characters over longer periods and, we imagine, are more likely to include scenes with normal, everyday uses of passwords. The relative normalcy of life on television (compared to in movies) might make these portrayals seem more realistic to viewers. Our work did not provide insights on *how* to expand the scope of analysis beyond the topic of passwords; however, our findings demonstrate that doing so could help solidify our knowledge about how cybersecurity and privacy topics (or technology more broadly) are portrayed in movies and, thus, what viewers might be learning. Along these lines, it was beyond the scope of our study to determine how the elements of movies we identified actually impact people, but this is an important next step. Finally, while we have suggested above that instructors could incorporate movie clips in their classes, other fields where this is common have conducted studies to understand how this should best be done and how to avoid common pitfalls; this type of follow-on work would be beneficial in this domain as well.

8   Conclusion

To analyze the depiction of passwords and password behavior in movies, we performed a subtitle analysis and watched selected scenes. Our results show a broad spectrum of different password activities and contexts of usage in movies from various years and genres. Movies show passwords of different strengths and outline different kinds of password attacks. However, the chances of success are presented as dangerously high and important best practices are missing from the portrayal. We aim to contribute towards a better understanding of how cybersecurity is depicted in the media, and ultimately to a better understanding of how we can mitigate the (negative) consequences of wrongful depiction of cybersecurity.

References

[1] Ruba Abu-Salma, M. Angela Sasse, Joseph Bonneau, Anastasia Danilova, Alena Naiakshina, and Matthew Smith. Obstacles to the Adoption of Secure Communication Tools. In *IEEE Symposium on Security and Privacy*, SP '17, pages 137–153, San Jose, California, USA, May 2017. IEEE.

[2] Yusuf Albayram, John Liu, and Stivi Cangonj. Comparing the Effectiveness of Text-based and Video-based Delivery in Motivating Users to Adopt a Password Manager. In *European Workshop on Usable Security*, EuroUSEC '21, pages 89–104, Virtual Conference, October 2021. ACM.

[3] Anonymous Reddit User. r/DataHoarder: 5,719,123 Subtitles From OpenSubtitles.org, July 2022. https://www.reddit.com/r/DataHoarder/comments/w7sgcz/5719123_subtitles_from_opensubtitlesorg/, as of June 6, 2024.

[4] Khadija Baig, Elisa Kazan, Kalpana Hundlani, Sana Maqsood, and Sonia Chiasson. Replication: Effects of Media on the Mental Models of Technical Users. In *Symposium on Usable Privacy and Security*, SOUPS '21, pages 119–138, Virtual Conference, August 2021. USENIX.

[5] Travis Bell and Community. The Movie Database: Popularity & Trending, November 2023. https://developer.themoviedb.org/docs/popularity-and-trending, as of June 6, 2024.

[6] Pablo González Blasco. Literature and Movies for Medical Students. *Family Medicine*, 33(6):426–428, June 2001.

[7] BonaFideBoss. IMDb: Top 1000 Highest-Grossing Movies of All Time, April 2024. https://www.imdb.com/list/ls098063263/, as of June 6, 2024.

[8] "Brano" and Community. Open Subtitles: Download Movie and TV Series Subtitles, January 2006. https://www.opensubtitles.org, as of June 6, 2024.

[9] Ulla Bunz. "We speak in code, in case the telephone operator should be eavesdropping!": How Popular Movies Reflect Society's Attitude Toward Technology. In *Annual Convention of the Media Ecology Association*, MEA '03, pages 1–18, Hempstead, New York, USA, June 2003. MEA.

[10] Andrew C. Butler, Franklin M. Zaromb, Keith B. Lyle, and Henry L. Roediger. Using Popular Films to Enhance Classroom Learning: The Good, the Bad, and the Interesting. *Psychological Science*, 20(9):1161–1168, September 2009.

[11] Sonia Chiasson and Paul C. Van Oorschot. Quantifying the Security Advantage of Password Expiration Policies. *Designs, Codes and Cryptography*, 77(2–3):401–408, December 2015.

[12] Mathieu Christmann, Peter Mayer, and Melanie Volkamer. Vision: What Johnny learns about Password Security from Videos posted on YouTube. In *European Workshop on Usable Security*, EuroUSEC '21, pages 124–128, Virtual Conference, October 2021. ACM.

[13] Wikipedia Community. Wikipedia: Film analysis, May 2024. https://en.wikipedia.org/w/index.php?title=Film_analysis&oldid=1213215167, as of June 6, 2024.

[14] Nik Cubrilovic. RockYou Hack: From Bad To Worse, December 2009. https://techcrunch.com/2009/12/14/rockyou-hack-security-myspace-facebook-passwords/, as of June 6, 2024.

[15] Maria Dimitropoulou, Jon Andoni Duñabeitia, Alberto Avilés, José Corral, and Manuel Carreiras. Subtitle-Based Word Frequencies as the Best Estimate of Reading Behavior: The Case of Greek. *Frontiers in Psychology*, 1:218:1–218:12, December 2010.

[16] "Fast Eddie" and Community. TV Tropes Pop-Culture Wiki: Hollywood Hacking, November 2010. https://tvtropes.org/pmwiki/pmwiki.php/Main/HollywoodHacking, as of June 6, 2024.

[17] "Fast Eddie" and Community. TV Tropes Pop-Culture Wiki: Hollywood Encryption, January 2014. https://tvtropes.org/pmwiki/pmwiki.php/Main/HollywoodEncryption, as of June 6, 2024.

[18] Stefan Engert and Alexander Spencer. International Relations at the Movies: Teaching and Learning about International Politics through Film. *Perspectives: Review of International Affairs*, 17(1):83–103, July 2009.

[19] Franklin Fearing. Influence of the Movies on Attitudes and Behavior. *The Annals of the American Academy of Political and Social Science*, 254:70–79, November 1947.

[20] Kelsey R. Fulton, Rebecca Gelles, Alexandra McKay, Yasmin Abdi, Richard Roberts, and Michelle L. Mazurek. The Effect of Entertainment Media on Mental Models of Computer Security. In *Symposium on Usable Privacy and Security*, SOUPS '19, pages 79–95, Santa Clara, California, USA, August 2019. USENIX.

[21] Brian Gallagher. 10 Great Social Commentary Movies That Reflect Contemporary Society, August 2017. https://www.tasteofcinema.com/2017/10-great-social-commentary-movies-that-reflect-contemporary-society/, as of June 6, 2024.

[22] Maximilian Golla and Markus Dürmuth. On the Accuracy of Password Strength Meters. In *ACM Conference on Computer and Communications Security*, CCS '18, pages 1567–1582, Toronto, Ontario, Canada, October 2018. ACM.

[23] Damian Gordon. Forty Years of Movie Hacking: Considering the Potential Implications of the Popular Media Representation of Computer Hackers from 1968 to 2008. *International Journal of Internet Technology and Secured Transactions*, 2(1/2):59–87, February 2010.

[24] Hana Habib, Pardis Emami Naeini, Summer Devlin, Maggie Oates, Chelse Swoopes, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. User Behaviors and Attitudes Under Password Expiration Policies. In *Symposium on Usable Privacy and Security*, SOUPS '18, pages 13–30, Baltimore, Maryland, USA, August 2018. USENIX.

[25] Ata Hakçıl ("ignis sec"). PWDB: New Generation of Password Mass-Analysis, July 2020. https://github.com/ignis-sec/Pwdb-Public, as of June 6, 2024.

[26] Cormac Herley and Paul C. Van Oorschot. A Research Agenda Acknowledging the Persistence of Passwords. *IEEE Security & Privacy*, 10(1):28–36, January 2012.

[27] Beth L. Hoffman, Ariel Shensa, Charles Wessel, Robert Hoffman, and Brian A. Primack. Exposure to Fictional Medical Television and Health: A Systematic Review. *Health Education Research*, 32(2):107–123, April 2017.

[28] Lee Humphreys, Phillipa Gill, and Balachander Krishnamurthy. Twitter: A Content Analysis of Personal Information. *Information, Communication & Society*, 17(7):843–857, October 2013.

[29] Friska Sari Luksiana Hutajulu and Herman Herman. Analysis of Illocutionary Act in the Movie "You Are My Home" English Subtitle. *Journal of English Educational Study*, 2(1):29–36, May 2019.

[30] Philip Nicholas Johnson-Laird, Vittorio Girotto, and Paolo Legrenzi. Mental Models: A Gentle Guide for Outsiders, April 1998. `https://web.archive.org/web/20050305184203/https://www.si.umich.edu/ICOS/gentleintro.html`, as of June 6, 2024.

[31] Aikaterini Kanta, Iwen Coisel, and Mark Scanlon. A Novel Dictionary Generation Methodology for Contextual-Based Password Cracking. *IEEE Access*, 10:59178–59188, June 2022.

[32] Aikaterini Kanta, Iwen Coisel, and Mark Scanlon. Harder, Better, Faster, Stronger: Optimising the Performance of Context-Based Password Cracking Dictionaries. *Forensic Science International: Digital Investigation*, 44:301507:1–301507:9, March 2023.

[33] Aikaterini Kanta, Sein Coray, Iwen Coisel, and Mark Scanlon. How Viable Is Password Cracking in Digital Forensic Investigation? Analyzing the Guessability of over 3.9 Billion Real-World Accounts. *Forensic Science International: Digital Investigation*, 37:301186:1–301186:11, July 2021.

[34] Gary Koch and Colette T. Dollarhide. Using a Popular Film in Counselor Education: Good Will Hunting as a Teaching Tool. *Counselor Education and Supervision*, 39(3):203–210, March 2000.

[35] Herbert E. Krugman and Eugene L. Hartley. Passive Learning From Television. *The Public Opinion Quarterly*, 34(2):184–190, June 1970.

[36] Marcus Law, Wilson Kwong, Farah Friesen, Paula Veinot, and Stella L. Ng. The Current Landscape of Television and Movies in Medical Education. *Perspectives on Medical Education*, 4(5):218–224, September 2015.

[37] Pierre Lison and Jörg Tiedemann. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. *10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 923–929, 2016. `https://opus.nlpl.eu/OpenSubtitles/corpus/version/OpenSubtitles`, as of June 6, 2024.

[38] Johnny Long ("j0hnny"). DEFCON 14: Secrets of the Hollywood Hacker!, August 2006. `https://www.youtube.com/watch?v=m_Xmc49ZrYA`, as of June 6, 2024.

[39] Abolfaz Mahdiloo and Siros Izadpanah. The Impact of Humorous Movie Clips on Better Learning of English Language Vocabulary. *International Journal of Research in English Education*, 2(2):16–30, June 2017.

[40] Paweł Mandera, Emmanuel Keuleers, Zofia Wodniecka, and Marc Brysbaert. Subtlex-Pl: Subtitle-Based Word Frequency Estimates for Polish. *Behavior Research Methods*, 47(2):471–483, June 2015.

[41] Peter Mayer, Collins W. Munyendo, Michelle L. Mazurek, and Adam J. Aviv. Why Users (Don't) Use Password Managers at a Large Educational Institution. In *USENIX Security Symposium*, SSYM '22, pages 1849–1866, Boston, Massachusetts, USA, August 2022. USENIX.

[42] Peter Mayer and Melanie Volkamer. Addressing Misconceptions about Password Security Effectively. In *Workshop on Socio-Technical Aspects in Security and Trust*, STAST '17, pages 16–27, Orlando, Florida, USA, December 2017. ACM.

[43] Daniel Miessler and Community. SecLists: "200 Most Used Passwords", December 2023. `https://github.com/danielmiessler/SecLists/blob/master/Passwords/2023-200_most_used_passwords.txt`, as of June 6, 2024.

[44] Hani Morgan. Counteracting Misconceptions about the Arab World from the Popular Media with Culturally-Authentic Teaching. *International Social Studies*, 2(2):70–83, January 2013.

[45] National Cyber Security Centre. The Problems with Forcing Regular Password Expiry, December 2016. `https://www.ncsc.gov.uk/articles/problems-forcing-regular-password-expiry`, as of June 6, 2024.

[46] Katharina Pfeffer, Alexandra Mai, Edgar Weippl, Emilee Rader, and Katharina Krombholz. Replication: Stories as Informal Lessons about Security. In *Symposium on Usable Privacy and Security*, SOUPS '22, pages 1–18, Boston, Massachusetts, USA, August 2022. USENIX.

[47] McKenna Princing. I Was a Medical Advisor for Grey's Anatomy. Here's What I Learned., October 2017. `https://rightasrain.uwmedicine.org/well/stories/i-was-medical-advisor-greys-anatomy-heres-what-i-learned`, as of June 6, 2024.

[48] Emilee Rader, Rick Wash, and Brandon Brooks. Stories as Informal Lessons about Security. In *Symposium on Usable Privacy and Security*, SOUPS '12, pages 6:1–6:17, Washington, District of Columbia, USA, July 2012. ACM.

[49] Pritham Y. Raj. Medicine, Myths, and the Movies. Hollywood's Misleading Depictions Affect Physicians, Patients Alike. *Postgraduate Medicine*, 113(6):9–10, June 2003.

[50] Mudassar Raza, Muhammad Iqbal, Muhammad Sharif, and Waqas Haider. A survey of password attacks and comparative analysis on methods for secure authentication. *World applied sciences journal*, 19(4):439–444, 2012.

[51] Elissa M. Redmiles, Sean Kross, and Michelle L. Mazurek. How I Learned to Be Secure: A Census-Representative Survey of Security Advice Sources and Behavior. In *ACM Conference on Computer and Communications Security*, CCS '16, pages 666–677, Vienna, Austria, October 2016. ACM.

[52] Elissa M. Redmiles, Amelia R. Malone, and Michelle L. Mazurek. I Think They're Trying to Tell Me Something: Advice Sources and Selection for Digital Security. In *IEEE Symposium on Security and Privacy*, SP '16, pages 272–288, Los Alamitos, CA, USA, May 2016. IEEE Computer Society.

[53] "Rhiannon". Hacker's Game: 10 Things Hollywood Got Wrong About Computer Hacking, July 2022. https://hotbotvpn.com/blog/10-things-hollywood-got-wrong-about-computer-hacking/, as of June 6, 2024.

[54] Scott Ruoti, Tyler Monson, Justin Wu, Daniel Zappala, and Kent Seamons. Weighing Context and Trade-offs: How Suburban Adults Selected Their Online Security Posture. In *Symposium on Usable Privacy and Security*, SOUPS '17, pages 211–228, Santa Clara, California, USA, July 2017. USENIX.

[55] Richard Shay, Saranga Komanduri, Adam L. Durity, Phillip (Seyoung) Huh, Michelle L. Mazurek, Sean M. Segreti, Blase Ur, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. Can long passwords be secure and usable? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 2927–2936, New York, NY, USA, April 2014. Association for Computing Machinery.

[56] University of North Carolina at Chapel Hill The Writing Center. Film analysis, May 2024. https://writingcenter.unc.edu/tips-and-tools/film-analysis/, as of June 6, 2024.

[57] Ghent University. DICIS – A Scientific Research Network on Digital Cinema Studies, May 2024. https://www.ugent.be/ps/communicatiewetenschappen/cims/en/research/current-research-projects/dicis.htm, as of June 6, 2024.

[58] Blase Ur, Jonathan Bees, Sean M. Segreti, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. Do Users' Perceptions of Password Security Match Reality? In *ACM Conference on Human Factors in Computing Systems*, CHI '16, pages 3748–3760, San Jose, California, USA, May 2016. ACM.

[59] Blase Ur, Sean M. Segreti, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Saranga Komanduri, Darya Kurilova, Michelle L. Mazurek, William Melicher, and Richard Shay. Measuring Real-World Accuracies and Biases in Modeling Password Guessability. In *USENIX Security Symposium*, SSYM '15, pages 463–481, Washington, District of Columbia, USA, August 2015. USENIX.

[60] USC Annenberg Norman Lear Center. Hollywood, Health and Society: About Us, February 2024. https://hollywoodhealthandsociety.org/about-us/, as of June 6, 2024.

[61] Niklas von Boguszewski, Sana Moin, Anirban Bhowmick, Seid Muhie Yimam, and Chris Biemann. How Hateful are Movies? A Study and Prediction on Movie Subtitles. *CoRR*, abs/2108.10724:1–12, August 2021.

[62] Ding Wang, Ping Wang, Debiao He, and Yuan Tian. Birthday, Name and Bifacial-Security: Understanding Passwords of Chinese Web Users. In *USENIX Security Symposium*, SSYM '19, pages 1537–1555, Santa Clara, California, USA, August 2019. USENIX.

[63] Rick Wash. Folk Models of Home Computer Security. In *Symposium on Usable Privacy and Security*, SOUPS '10, pages 11:1–11:16, Redmond, Washington, USA, July 2010. ACM.

[64] OpenSubtitles Webmasters. Disclaimer - opensubtitles.org, May 2024. https://www.opensubtitles.org/de/disclaimer, as of June 6, 2024.

[65] Danny Wedding and Ryan M. Niemiec. *Movies and Mental Illness: Using Films to Understand Psychopathology*. Hogrefe Publishing, Göttingen, Germany, 3 edition, 2009.

[66] Daniel Lowe Wheeler. zxcvbn: Low-Budget Password Strength Estimation. In *USENIX Security Symposium*, SSYM '16, pages 157–173, Austin, Texas, USA, August 2016. USENIX.

[67] Naomi White. How Hollywood Movies Reflect Society, September 2022. https://www.thegreatdebatersmovie.com/how-hollywood-movies-reflect-society/, as of June 6, 2024.

[68] David Knapp Whittier, May G. Kennedy, Janet S. St. Lawrence, Salvatore Seeley, and Vicki Beck. Embedding Health Messages into Entertainment Television: Effect on Gay Men's Response to a Syphilis Outbreak. *Journal of Health Communication*, 10(3):251–259, September 2005.

[69] Tannis Macbeth Williams. How and What Do Children Learn from Television? *Human Communication Research*, 7(2):180–192, December 1981.

[70] Muheng Yu, Michael C. Carter, Drew P. Cingel, and Jeanette B. Ruiz. A Content Analysis of Aggression in Netflix Original, Adolescent-Directed Series' Subtitles. *Communication Quarterly*, 71(5):588–609, August 2023.

[71] Muheng Yu, Michael C. Carter, Drew P. Cingel, and Jeanette B. Ruiz. How Sex Is Referenced in Netflix Original, Adolescent-Directed Series: A Content Analysis of Subtitles. *Psychology of Popular Media*, 13(1):1–11, January 2024.

[72] Samira Zibaei, Dinah Rinoa Malapaya, Benjamin Mercier, Amirali Salehi-Abari, and Julie Thorpe. Do Password Managers Nudge Secure (Random) Passwords? In *Symposium on Usable Privacy and Security*, SOUPS '22, pages 581–597, Boston, MA, USA, August 2022. USENIX.

[73] Verena Zimmermann and Nina Gerber. "If It Wasn't Secure, They Would Not Use It in the Movies" - Security Perceptions and User Acceptance of Authentication Technologies. In *Human Aspects of Information Security, Privacy and Trust*, HAS '17, pages 265–283, Vancouver, British Columbia, Canada, July 2017. Springer.

A Passwords in Movies: List and Strength Analysis Results with *zxcvbn*

Passwords were repeatedly seen in the analyzed movies. With zxcvbn these are sorted into five Strength Categories. Below we show excerpts from the list. The complete list can be viewed online[3].

---

[3]https://www.itsec.uni-hannover.de/de/usec/forschung/medien/password-depiction-in-movies

Class 0 (133) 0000, 01234, 0515, 1111, 1212, 123, 123123, 123321, 1234, 12345, 123456, 164, 179, 1951, 1967, 1972, 1982, 1998, 2222, 2345, 2468, 286, 314, 314159, 326, 38, 4040, 4321, 437, 438, 500, 521, 651, 680, 69, 696969, 761, 77777, 923, 949, 999999, 999999999, a2h, ABC123, ABCD1234, Angela, Anna, Annie, Barbara, Batman, beer, Birdie, Bob, Boobs, butterfly, carmen, Casper, Crystal, Denise, diamond, Die, eat, Enter, Eric, Erica, erin, Faye, Frankie, freedom, girls, guess, guest, h0us3, Heaven, Horny, James, Jenny, Justin, Laura, leon, Love, Lucas, March, Melanie, Mountain, Myself, Natasha, nose, Om, Orlando, Paradise, Party, password, Password, PASSWORD, Peaches, Pedro, Pepper, pirate, Porn, princess

Class 1 (317) 0113, 040515, 0511, 0512, 05171210, 0522, 0623, 0627, 070476, 0708, 0710, 072099, 0801, 1048, 1104, 1112, 1126, 1166, 1192, 1195, 1230, 1321, 132109, 1356, 1492, 15626, 1685198, 1776, 1796, 19300830, 19891023, 20107, 20131026, 2111, 22093, 2235, 2259, 2356, 2372, 2501, 2598, 262670, 27130, 295141, 2QUILA, 3041, 3057, 3690, 4093, 420God, 4664, 489*48, 4989, 5023, 5042, 5321, 541267, 5445, 54AGT, 5urvivor, 6143, 6246, 627628, 661968, 691234, 712735, 7232, 7397, 7590, 8224, 8644, 8854, 977127, 9993, a/;321, A3501, Abby, Abdel, adventure, AirBud, Amen, AngeIo, Angelangel, Angiovanni, annie123, Anusua, argonaut, athlete, Autumnleave, Bacon, badmama, baloney1, Barnsey, Bassola, Bastard, beagles, BEARD, Beethoven, Begood, Belle1998, BigBen, Biggie, BlackChicken, BlackOut, bluebeauty, Bondik, Boomerang, boxing, BRANGELINA, Brat, Briefs, Brigitte, Buremma, buttercup, Cancerian, Carmim, casket, catnip, catnip1, chandelier, Charmer, Chestnut, chicken65, Chicken65, ChiefAsshole, ChowChow, chrisnewton, CM110

Class 2 (118) 0505informer, 1/2-1/2, 14-J-89, 2060Pinto, 2516904, 801023000000, ACAPULCO01, asavari, asstastic, ATR1020, auroraborealis, AVCHomes, Ayla123, B055man69, badmamma, Balki1987, BaluMama, Bankerchick, Batfan1, bayernmunich, Beatrix928, BettyGrable, BODYGARD, canttell, Carlton071133, CarryGold, Césoul89, Charbear, Chewinggum, Clavius, Cloudberry, Creamcrackers, cuddlefresh, Cv'qrPo, Damnedmelon, darlinggoli, ddayspm, DEADRIPLEY, DeathWhisper, DevAnand, DJDESFAS, Doorlogs, ExtraStuffing, Fartnoise, gindrick, Haircomb, HAPPYMANPAN, HDA14+1, Helvetica, hoodfume, HumphreyBogart, IAN&EMMA, ilikelaura, imthman, interzone

Class 3 (70) 13C34RMXL, 2015salesstrategy, 45gx67kn21, 4saraandjimmy, 68k305RW65, Abhimanyu, Abraxas79713, arthurisadick, asami0709, Barbsguy1989, bardahlia13, batchoy0706, bauer-smythe, bethmarch4eva, Blumenfeld, boobfart69, broccoli34525A, ch3rryjone3s, cocknballs, daddysprincess1994, DavidFosterWallace, DirtyDaniels69, dividebyzero, Dongmaster82, Donkeyballs84, Effenberg, Eightclap1, Elchapo69, fluxcapacitor, GOD'SGIFT, Grid90245, Grilledcarrots, Hananamiti, Hasselhof, Heaven'sDoor, icanttellyou, ILoveYudi, Indiansubcontinent79, Johnnyutah69, Konigshutte, liz0919/85, louisepaul222, MaRc62?!*$, meganthemoron, milkandcookie$, MillerEmployeeGeneral, minayo0118, misterdarcyforever

Class 4 (49) 72435637440472, Aatukaalamma, alfreddabuttler, AndrétheGiant, arianagrandespuffynipples555, Asagolabius, Bagofdicks44, Baitursinov, Bergen-Belsen, BickmanGuest, bigbertandsmallbert, bryansbabydick69, CafeBonaparte, Cantarpiano1863, catwomanisaBitch, DabanggSultan, DOOMEDIFYQUQUIT, Epluribusfunk, EvianBottledAir, GabriMarta202, HAWTHRONE1850, Heymonaumona, ILOVEchotaBHEEM143, itsagratefuldead, JoyMukherji, k!TTenT!ckler312, Kavya_Kavya, ldfvarumellamsheriaavum, Marsupilami, MeikoMochizuki, MilkyShonku21, ninelformiguel75, OffWithTheirHeads!, penis_grigio72, piazzadellecinquelune, Prabhavathi, ROSEPOGONIAS, StarBigSkyChristmas!, Stephanie'sdude2016

B   Points in the Password Life Cycle: Descriptions and Sample Scenes

During the password life cycle different activities are performed, as described in Table 1.

Life cycle point	Description	Example
Password Creation	Setting up a password or speaking about password generation	"He said that he would create a website. In order to access the website, I would need a password. So he took a paper napkin that was on the table in this cafe where we were talking in Brussels and he hooked together several of the words in the commercial logo [...]" (We Steal Secrets: The Story of WikiLeaks, 2013)
Password Change	Changing the password as account owner or legitimated person or intending to	"To flush out the mole is easy, change our password and signals, tell all the others, and pretend nothing's wrong." (The Swordswoman in White, 1992)
Training to Remember a Password	Checking if someone else still remembers the password or reminding them of it	"I'll see you in an hour? -Right. -Haven't forgotten the password? -Whatever gave you the idea?" (The Body, 2003)
Password Loss and Oblivion	Failing to remember the password or losing a physical reminder like a piece of paper	"My favorite is when they come in, forgotten their password, Locked themself out of their own computer." (The Zombie Werewolves Attack!, 2009)
Password Recovery, Reset, and Hints	Changing the password using recovery systems or receiving hints to remember the password	"All that you do is enter an email address and attempt to enter a password. Then, you see, it asks if you forgot your password. So you click that and it tells you to check your email to change your password. So then I go to her email [...] (16 and Missing, 2015)
Password Reuse	Reusing the same password for multiple different purposes or accounts	"will need the passwords to your email accounts, your social media accounts, your bank accounts, your credit card accounts and your Cinnabon Rewards account. - It's easy. It's the same password for all of 'em. It's phil123456. - You've got to be kidding me." (Jexi, 2019)

Table 1: Password activities within the movies: Activity names, descriptions, and example scenes.

C   Comparison of the Dataset with High-Grossing Movies

High-grossing movies containing the word "password" (70):   In the *Top 1000 Highest-Grossing Movies of all Time* [7] we identified 70 movies containing the word *password* at least once (see Section 3). Those are:

Alice in Wonderland; Armageddon; Avengers: Endgame; Batman Returns; Bruce Almighty; Captain America: The Winter Soldier; Captain Marvel; Captain Phillips; Casino Royale; Cheaper by the Dozen; Crazy Alien; Crazy Rich Asians; Disclosure; Doctor Strange; Elysium; Ghost; Godzilla vs. Kong; GoldenEye; Harry Potter and the Chamber of Secrets; Harry Potter and the Prisoner of Azkaban; Harry Potter and the Sorcerer's Stone; Heat; Home; Ice Age: Continental Drift; Iron Man 3; It; It Chapter Two; Kingsman: The Secret Service; Lucy; Men in Black: III; Mojin: The Lost Legend; Monster Hunt; National Treasure; Ne Zha; Non-Stop; Now You See Me; Parasite; Pitch Perfect 2; Ralph Breaks the Internet; Ready Player One; Safe House; Sex and the City; Spider-Man: Far from Home; Spider-Man: Into the Spider-Verse; Superman Returns; Tangled; Terminator Genisys; The Batman; The Bodyguard; The Break-Up; The Departed; The Emoji Movie; The Firm; The Hangover Part II; The Hangover Part III; The Hitman's Bodyguard; The Incredibles; The Intern; The Lego Batman Movie; The Lord of the Rings: The Fellowship of the Ring; The Other Woman; The Pacifier; The Secret Life of Pets; The Shape of Water; The Social Network; The SpongeBob Movie: Sponge Out of Water; The Vow; Tomorrowland; True Lies; Who Framed Roger Rabbit

Comparative Statistics: For each topic of the paper, the code frequencies were calculated for the high-grossing movies and the entire data set. The distribution within the different topics is compared in Figure 11 and in Figure 12 the differences in attacks are shown.
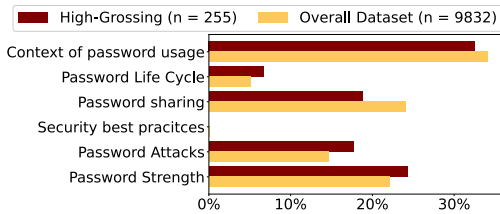
(a) **Overview of topic distribution** This indicator of which topics occur rather frequently or rarely within the data set shows, that basically, all topics except best practices are present in both data sets and also with approximately similar distribution.
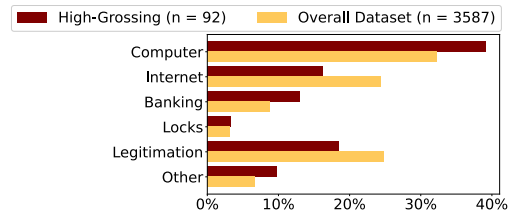
(b) **Context of Use** As described in Section 4.2, it is in often identifiable what the password is used for. In the high-grossing movies there are more movies used in the compute context and less regarding Internet and Legitimation.
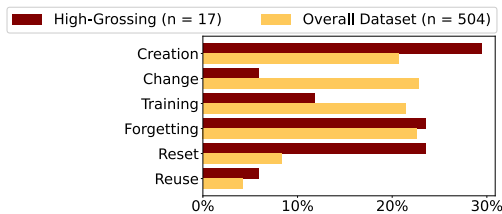
(c) **Life Cycle** The different scenarios in the life cycle of passwords (cf. Section 4.2) occur in the High-Grossing Movies but with different characteristics. Only one movie contains Change and Reuse, only two Training.
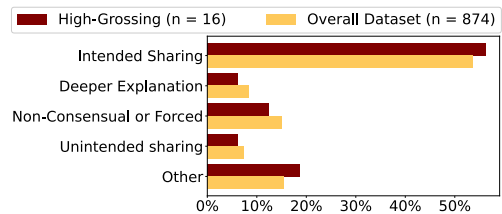
(d) **Password sharing** In 16 scenes in the top-grossing movies password sharing is presented. The frequency distribution is very similar to that of the full dataset. Similar as in the whole dataset, most common type is *Intended/Intentional* sharing.
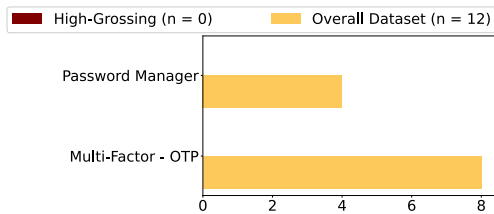
(e) **Security Best Practices in total numbers** This topic appears only in the overall dataset and not in the top-grossing movies. The number of scenes (12) is negligible (0.12% of codes in the set of all codes applied to the overall dataset).
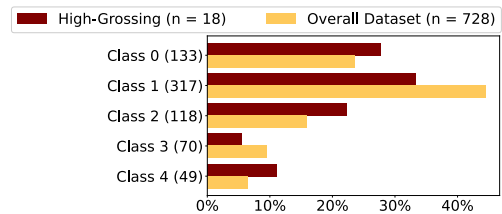
(f) **Password Strength: zxcvbn** Passwords with all strength categories using zxcvbn appear in both the high-grossing and the overall dataset. High-grossing movies contain fewer class 1 passwords than the overall dataset.

Figure 11: Comparison of high-grossing and overall dataset for different topics presented in this paper. The number of codes per topic was calculated for each diagram for both the high-grossing and the overall dataset and was used to quote the distribution.
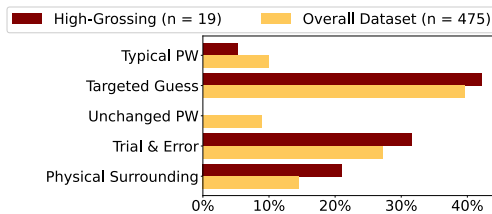
(a) **Password Guessing: Approach**   In the top-grossing movies, unlike in the overall dataset, no unchanged password is used for guessing. Otherwise, the frequency of the procedures is similar.
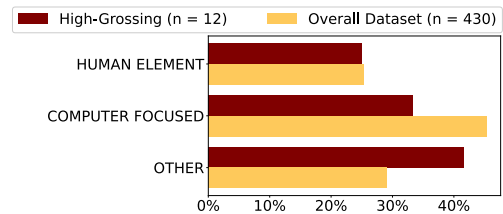
(b) **Password Hacking: Approach**   There are only 12 scenes, therefore rough categories are used. The high-grossing movies contain more "other" and less computer-based approaches.
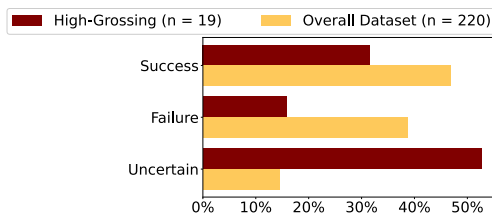
(c) **Password Guessing: Success**   For the high-grossing movies, an above-average number of the guessing approaches have an unclear outcome. These are relatively rare in the overall dataset. Similar to the overall dataset, more attacks are successful than unsuccessful.

(d) **Password Hacking: Success**   For the High-grossing movies, exactly half of the hacking attempts are successful and the other half have an unclear outcome. Not a single attempt fails. This is very different from the overall dataset, where most attacks are successful.

(e) **Password Guessing: Success and Strength Category**   Only five scenes with password guessing and a shown password (allowing strength class analysis). Four times this is a class 0 password, one time a Class 1.

(f) **Password Hacking: Success and Strength Category**   Not a single scene with password hacking and a shown password (allowing strength class analysis). Accordingly, no further analysis is possible.

Figure 12: Comparison of high-grossing movies and movie-dataset for password attacks. The number of codes per topic was calculated for each diagram for both the high-grossing and the overall dataset. This was used to calculate the percentages.

Understanding How People Share Passwords

Phoebe Moh, Andrew Yang, Nathan Malkin[*], Michelle L. Mazurek
University of Maryland, [*]New Jersey Institute of Technology

Abstract

Many systems are built around the assumption that one account corresponds to one user. Likewise, password creation and management is often studied in the context of single-user accounts. However, account and credential sharing is commonplace, and password generation has not been thoroughly investigated in accounts shared among multiple users. We examine account sharing behaviors, as well as strategies and motivations for creating shared passwords, through a census-representative survey of U.S. users ($n = 300$). We found that password creation for shared accounts tends to be an individual, rather than collaborative, process. While users tend to have broadly similar password creation strategies and goals for both their personal and shared accounts, they sometimes make security concessions in order to improve password usability and account accessibility in shared accounts. Password reuse is common among accounts collectively shared within a group, and almost a third of our participants either directly reuse or reuse a variant of a personal account password on a shared account. Based on our findings, we make recommendations for developers to facilitate safe sharing practices.

1 Introduction

It is generally assumed that an individual's password is a secret that no one else knows; yet, in reality, sharing passwords for online accounts is widespread. People share credentials for a variety of rational reasons, including for work, finances, convenience, or as a sign of trust among romantic partners and family members [4, 32, 33, 34, 43]. Others share accounts

with trusted parties out of necessity, such as in the case of refugees, older adults, and other members of at-risk populations [26, 42, 49]. In many such cases, users perceive these needs to be a higher priority than account security.

Instead of repeatedly and ineffectually warning users against sharing [52], technology creators and security experts should endeavor to design systems that take into account the reality of sharing. To do so effectively, it is important to understand how credential sharing works in practice. How users create and distribute passwords for shared accounts has important security implications. If users judiciously create unique credentials for sharing, then perhaps the current emphasis on discouraging sharing is misplaced. On the other hand, reusing passwords across personal and shared accounts creates risks to these personal accounts. In this case, new interventions—whether in terms of new system designs that accommodate sharing, better user education, or both—may be needed.

Further, while password creation has been extensively studied in the single-user setting [7, 46], less attention has been devoted to shared accounts. If password creation strategies for intended-to-be-shared accounts differ importantly from single-user-account strategies, different guidance (password meters, strength requirements, suggested strong passwords) may be needed. If these passwords are created collaboratively with the input of multiple users, rather than being dictated by a single user, the situation may be even more complex.

Thus, by understanding how and by whom shared account passwords are created, the motivations behind password strategies these users employ, and whether these shared passwords are reused and in what context(s), we hope to inform safe account sharing practices and design. To do this, we study the following research questions:

RQ1: **Is password creation in shared accounts a collaborative process, or is it predominantly individual? Who is involved in the password creation process?**

RQ2: **When users create passwords for shared accounts, are their priorities and strategies similar to when**

they create passwords for personal (non-shared) accounts?

RQ3: **How prevalent is password reuse among shared accounts? Are these passwords also reused for personal accounts?**

To answer these questions, we conducted a census-representative online survey ($n = 300$) among U.S. users. We found that participants tend to share accounts (predominantly streaming accounts) with a small number of users, typically romantic partners and family members. In addition:

- Approximately half of accounts surveyed were originally created with the intention of being shared; the other half began as personal accounts that later became shared. In the latter case, users often do not change the passwords of these personal accounts when they begin to share them.

- Password creation for shared accounts tends to be an individual rather than a collaborative process, and users tend to have similar password creation strategies for both their personal and shared accounts. However, password makers will sometimes take the capabilities of other users into consideration or make security concessions in order to improve access to the shared account.

- Shared-account passwords are frequently reused. Users often have a *group password* for sharing multiple accounts among approximately the same set of users. More concerningly from a security point of view, about a third of participants report reusing (either exact or variant-of) passwords between personal and shared accounts.

Based on our findings, we provide recommendations for technology creators to facilitate safe account sharing while minimizing potential harms.

2 Related work

Reasons and contexts for credential sharing. Account sharing within households and among romantic partners is driven by convenience, practicality, and reinforcing trust [24, 34]. Similarly, account sharing can be used to affirmation of trust between adolescents [33]. Customers of paid accounts have a financial incentive for sharing [18].

People may also share credentials out of necessity. Members of at-risk populations, such as refugees or older adults, often rely on trusted parties for important tasks or to maintain safety [26, 42, 49]. For example, Kenyan cybercafe customers with limited experience with computers sometimes rely on cybercafe managers to remember and manage their account login credentials in order to access essential services [27].

In workplace settings, coworkers share credentials to facilitate sharing files and resources [43], though difficulties often arise from working around systems built on the one user,

one account assumption [16, 21]. Cultural norms and expectations can be another driving reason for credential sharing, e.g., among bank customers in Saudi Arabia [4] or young adults in Bangladesh [3].

Account sharing can also continue after people want to stop it. In particular, Park et al. highlighted the difficulties of ending account sharing for romantic partners in the event of breakups, and Obada-Obieh et al. examined the cognitive and psychosocial burdens associated with ending account sharing [30, 34]. While we do not focus on adversarial relationships in this work, account sharing can also be used as a means of surveillance [5].

Kaye argued that password sharing is a nuanced social process rather than a deviant behavior to eliminate [19]. Indeed, these complex social processes are often important for maintaining security—for example, by small groups sharing digital resources to defend against insider and outsider threats [51, 53]. Some paradigms, such as family accounts, embrace account sharing and are designed around allowing multiple users to use a single account in an effort to enforce security without relying solely on social norms [12].

Taking into account the multitude of reasons for account sharing, password sharing is not likely to disappear anytime soon. Although motivations for account sharing are well-documented, the next step in the process—creating a password for the shared account—is not. Our study addresses this gap in knowledge.

Password generation and management by individual users. How people choose passwords for single-user accounts has long been studied both in the field and in the lab. Bryant and Campbell found that their surveyed participants often used meaningful data, such as nicknames, in their email passwords, and both partial and exact reuse of passwords across accounts was common [7]. Ur et al. observed password creation in the lab, finding that while most participants had a well-defined process for creating new passwords, many had misconceptions of what makes a password secure [46]. Studies comparing the security behaviors of experts with non-experts have found that non-experts tend to rely on memory to recall their passwords [8, 17].

Despite attempts at educating users, insecure behavior around passwords persists, often stemming from users' attempts to cope with the sheer number of passwords in daily life. In Stobert and Biddle's 2014 interview study, participants reported having a median of 27 accounts, and the authors found that these users ration effort to best protect important accounts by adopting less secure behaviors, such as reuse and writing down passwords, for accounts they deemed less sensitive or less frequently used [44]. Ur et al. and von Zezschwitz et al. similarly observed participants constructing weaker passwords for less-sensitive accounts [46, 48].

Partial or full reuse of old passwords represents one common effort rationing strategy. Das et al. estimated in 2014 that

43% of their participants directly reused passwords [10]. Shay et al. observed that most of their participants opted to modify an old password instead of creating an entirely new one in response to a university's password policy change [41]. Inglesant and Sasse noted that their participants often used "good" passwords as a resource to generate new passwords, and von Zezschwitz et al. found that weak passwords used by interviewed participants had roots in the first passwords they created [16, 48]. Wash et al. observed participants reusing passwords that were more complex and frequently-entered [50]. Users in Hanamsagar et al.'s study willingly traded security for memorability by reusing passwords in order to manage having many accounts [15]. Misconceptions about the risk of attacks and attacker capabilities were also a contributing factor to password reuse and weak passwords [15, 47]. Often, modifications made to old passwords to generate new passwords are small enough for an attacker aware of typical user behavior to guess the new password [10, 54].

While password creation strategies and motivations have been well-studied in the single-user context, we seek to expand this understanding to the multi-user context. Our explicit focus on shared accounts is an important lens for considering password behaviors employed by users.

Password managers. Security experts often recommend password managers as a means for users to cope with the ever-increasing number of passwords [6]. However, despite password managers' utility, only a relatively small proportion of users employ them. Those who do not use password managers, such as older adults, often cite security concerns and a lack of trust in password managers [35, 36]. While some password managers have aimed to support multi-user contexts [1, 22, 29], it remains unclear how often these features are used.

3 Methods

We designed our survey to understand how people share passwords in their day-to-day lives. We initially developed our protocol by adapting questions from related work on single-user password creation and interviewing seven people in the researchers' personal networks about their account sharing behaviors [7, 10, 40, 47]. To gather feedback, we piloted our survey with eight participants in think-aloud interviews and revised survey wording and presentation for clarity between interviews. Before final deployment, we further piloted the survey with 10 online participants.

Data collection took place in January and February 2023. Table 1 shows the demographic breakdown of our participants. All participants provided informed consent before beginning the survey, and the study was approved by the University of Maryland's institutional review board (IRB).

		Percent	Count
Gender	Female	47.7%	143
	Male	50.3%	151
	Nonbinary	<1.0%	1
Age	18-29	21.0%	63
	30-39	18.0%	54
	40-49	18.3%	55
	50-59	17.7%	53
	60+	23.3%	71
Annual household income	<$50k	33.7%	101
	$50k - $100k	38.7%	116
	>$100k	24.3%	73
Education	<High school	1.0%	3
	High school or equiv.	29.0%	87
	Bachelor or associate	52.0%	155
	Advanced degree	16.7%	50
CS background	Yes	19.3%	58
	No	78.7%	236
Security background	Yes	15.7%	47
	No	82.3%	247

Table 1: Participant demographics. Excludes "no answer" and "prefer not to say" options.

3.1 Survey protocol

Shared accounts (overview). Participants provided consent and then gave an overview of the accounts they shared. We defined a shared account as "any account where you and at least one other person both use the same username (or email address) and password combination in order to access and use the account, either at the same time or taking turns." (Accounts shared without any kind of password exchange were excluded.) For each shared account, respondents self-reported the service the account was for,[1] the type of account, and with whom they shared the account. Account type options presented to participants were initially derived from Park et al.'s survey on shared accounts in romantic relationships [34]. We derived additional account types, such as VPNs, from our own pilots. Table 2 shows the types of accounts participants reported sharing.

Personal accounts. For each shared account type (as defined in Table 2), we asked participants if they had any personal accounts (*not* shared with anyone else) of the same account type. For every account type the participant reported having both personal and shared accounts for, we asked the participant about the strategies and factors that influenced the creation of the password for one such personal account.

---

[1]Participants were not required to name the service explicitly. See *Survey Instrument* in the Appendix for wording details.

Account Type	Accounts	Participants	Examples
Video/Music Streaming	67.8% (665)	91.0% (273)	Netflix, Youtube, Hulu, Disney+, HBO Max, Apple Music
Shopping	14.7% (144)	42.0% (126)	Walmart, Amazon Prime, Newegg, Ticketmaster, Costco
Finances	5.2% (51)	10.7% (32)	Bank of America, Paypal, Chase, Mint, Fidelity
Rent/Utilities	4.7% (46)	9.3% (28)	Accounts for water, rent portals, mortgage accounts, Xfinity
Gaming	1.3% (13)	4.0% (12)	Steam, Xbox Live, Playstation Plus
File Sharing	1.3% (13)	3.7% (11)	iCloud, Google Drive, Box, Dropbox
Social Media	0.1% (9)	2.0% (6)	Instagram, Twitter, Facebook, Snap Chat
Productivity Tools	0.1% (6)	1.3% (4)	Google Calendar, Trello, Zoom, Canva
E-books	0.1% (6)	2.0% (6)	Kindle, Audible, Viz Media
News	0.1% (5)	1.7% (5)	New York Times, Consumer Reports, local newspapers
VPNs	<0.1% (4)	1.3% (4)	NordVPN, SurfShark
Health Insurance/Services	<0.1% (4)	1.0% (3)	Aetna, Cigna, OptumRx
Travel	<0.1% (4)	1.0% (3)	Websites for cruise lines and vacation rentals
E-mail	<0.1% (3)	1.0% (3)	Gmail, other e-mail services
Other	0.1% (8)	2.3% (7)	

Table 2: Types of accounts reported in the introduction of the survey (981 accounts total). For accounts that provide multiple services (such as Amazon Prime, which provides both shopping and streaming services), the account type was based on the dropdown option the participant selected.

Shared accounts (detailed). For the first four accounts the participant reported in the "shared accounts (overview)" section, we asked follow-up questions about the account, such as who was involved in password creation. If the participant was directly involved in password creation, we asked about the strategies they used and the motivations behind them. We chose to limit this section to the first four accounts reported (or all accounts, if fewer than four) in order to maximize recall and keep survey times manageable. We based the cutoff number on our pilots; pilot participants reported sharing an average of 3.6 accounts each. Because participants in our full study reported sharing an average of 3.3 accounts each, we believe that we achieved reasonable coverage.

Demographics. The survey concluded with demographic questions, which included income, education level, and background in computer/information security.

Data protection measures. We instructed participants not to share their passwords with us and periodically reminded participants that they should not enter their passwords into the survey. Furthermore, we did not collect any directly identifying from participants; participants were only identified by an anonymous Prolific (https://prolific.co) platform ID.

3.2 Recruitment

We recruited 300 respondents for our survey using Prolific's representative sampling feature, which recruits a demographically-representative (based on census data) sample of the U.S. population according to age, sex, and ethnicity (Table 1). We chose the sample size based on related survey

work [34, 41, 47, 50]. Participants were required to reside within the U.S., be at least 18 years old, and self-report fluency in English. The survey took an average of 16.5 minutes to complete (median 13.9 minutes), and participants were paid $3.75. We asked that participants have at least one account they shared with others in order to take the survey. If participants did not report any shared accounts, we discarded their responses (three overall). We used responses to open-ended questions to validate the quality of data collected, discarding responses (two overall) where participants provided off-topic answers. For discarded responses, we recruited new participants in their place to keep the final number of valid participants at 300.

3.3 Analysis

For open-ended answers, two coders collaboratively applied open-coding content analysis to draw out common themes around password creation and account sharing from responses, as well as surface themes that the researchers may not have initially been expecting. We used responses from the pilot studies and 10% of responses from the full survey to inductively develop an initial codebook [38]. Pilot responses were only used to develop the initial codebook, and we excluded pilot data from the final counts of codes and the remainder of the analysis. After creating the initial codebook, coders independently applied the codebook to an additional 10% of the responses from the full survey and met to discuss codes. Coders repeated this process three times, at which point code saturation and consensus was reached as measured by Cohen's kappa ($\kappa = 0.70$, indicating "substantial agreement") [20, 25, 38]. The remaining responses were divided among the coders to code independently. One coder reapplied
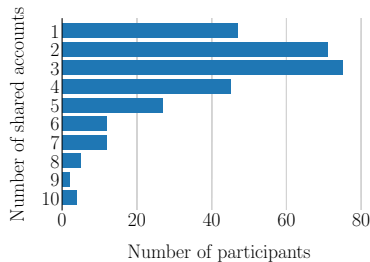
Figure 1: Number of accounts shared by each participant (981 accounts)
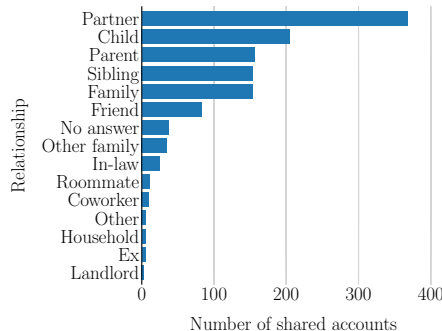
Figure 2: Who participants share their accounts with (981 accounts)
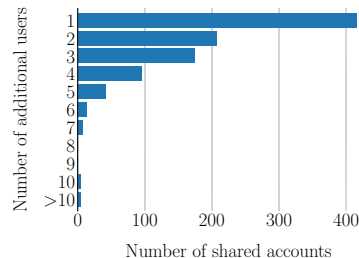
Figure 3: How many people (excluding themselves) participants share their accounts with (981 accounts)

the codebook to all prior responses from the codebook development phase to ensure that the final codes were adequately reflected across all responses.

In order to identify differences in password creation between personal and shared accounts, we built a regression model relating Likert-type responses about password-creation factors to whether an account was shared or personal. Details of this analysis are in Section 4.5.

3.4 Limitations

Our study has limitations inherent to online survey studies. Participants self-reported their password sharing behaviors, and we were unable to ask follow-up or clarification questions. Because we did not collect participants' passwords, we are unable to evaluate how secure these passwords actually were.

Due to social desirability and stigma against credential sharing, participants may not have reported all the accounts they share and may have underplayed insecure behaviors. However, some participants acknowledged their insecure practices ("I use the same password + variants for everything (bad — I know!)," Participant 70; "I know we aren't supposed to be reusing passwords, but this is the best one I have, and I can remember it," Participant 95; "I should have a more secure password, but I don't," Participant 177), and we believe that they were generally honest about their behaviors.

We asked participants in-depth questions about the first four accounts they reported rather than four random accounts in order to maximize recall, which may have biased which accounts were discussed for the participants who reported sharing more than four accounts.

We did not focus on adversarial account sharing or negative outcomes related to password sharing, and our participants did not discuss these topics in their free response answers. As such, our work only applies to voluntary account sharing.

Our sample size is not sufficient to obtain generalizable quantities for some uncommonly shared types of accounts (like VPN accounts). We focused on obtaining a broader view of the kinds of accounts people share rather than focusing on specific types of accounts.

Populations of online crowdsourcing platforms are generally more technologically-savvy than average; nonetheless, they provide reasonable sample populations [37]. In particular, Prolific has been found to be generally representative for user perceptions and experiences [45]. Furthermore, our usage of the platform's demographically-representative sampling feature ensured broader coverage of the U.S. population.

Because culture heavily influences expectations and norms surrounding credential sharing [3, 4, 27, 39], we focused on a single cultural context. Applying our research questions to non-U.S. contexts remains a subject for future research.

4 Results

We begin by describing our participants and the accounts they share (Section 4.1) and the prevalence of collaborative password-making for these shared accounts (Section 4.2). Next, we examine password reuse and other security behaviors (Sections 4.3 and 4.4), and finally we compare password creation strategies and motivations for shared accounts with those of personal accounts (Section 4.5).

4.1 Types of accounts participants share and with whom

Streaming accounts are the most common accounts shared by participants. Participants tend to share accounts with a few people close to them, typically partners and family members.

Our 300 participants reported sharing an average of 3.3 (median 3) accounts each. Figure 1 shows the distribution of shared accounts. Table 2 shows the types of accounts participants reported sharing; video and music streaming accounts are most popular, with 273 participants (91.0%) sharing at least one such account. Shopping accounts are the second most popular account to share, being shared by 126 participants (42.0%). Figure 2 shows with whom participants share
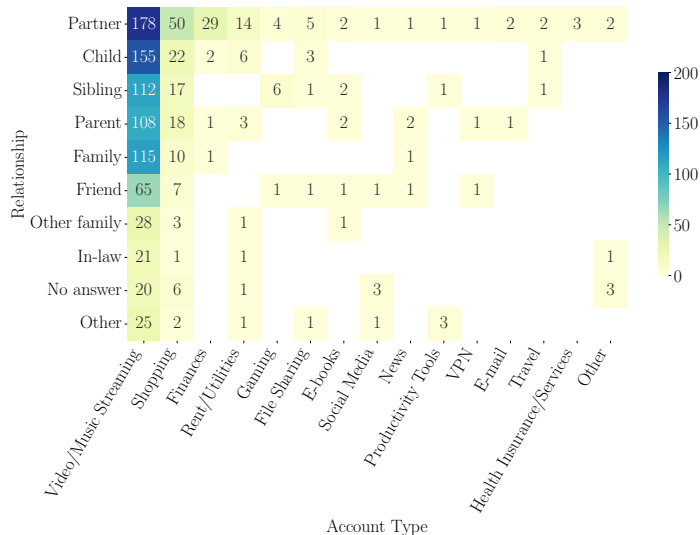
Figure 4: Sharing recipients, broken down by account type, from the "detailed" section of the survey (843 accounts total)
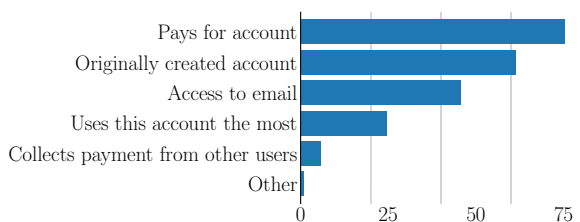
Figure 5: Factors contributing to account ownership. Participants could select more than one option per account.

their accounts. This is most commonly significant others and family members.

We report how many people each account was shared with in Figure 3. Accounts tend to be shared with a relatively small number of people: the median number of additional people (excluding the participant) an account was shared with was 2 (mean: 2.2 [2]). While not common, 20 participants (6.7%) report being unsure of exactly how many people have access to at least one of their shared accounts (35 accounts total).

After our 300 participants listed all of their shared accounts, we asked about the first four of them (or all of them, if fewer than four) in detail; the following results are based on these 843 shared accounts. Figure 4 summarizes who the subset of accounts we examine going forward are shared with and what type of accounts they are.[3]

---

[2]This statistic excludes 4 accounts for which participants selected the "> 10" option when reporting the number of other users, instead of specifying an exact count.

[3]This subset of accounts is representative of the broader set of shared accounts collected (cf. Figure 7 in the Appendix).

	Ownership			
Collaboration	**Single**	**Multi**	**Other**	*Total*
Single	632	115	0	*747*
Collaborative	9	29	1	*39*
Password generator	23	10	1	*34*
Other	5	5	13	*23*
Total	*669*	*159*	*15*	*843*

Table 3: Account ownership and collaboration in making shared passwords. *Other* includes unsure and no response.

4.2 Makers of shared passwords

Users rarely collaborate to make passwords for shared accounts. Generally, password creation is left to the sole discretion of a single account owner.

Shared accounts may have a single user who acts as the account owner (a single-owner shared account), or multiple users that share account ownership equally (a multi-owner shared account). We asked participants to identify their shared accounts as either single-owner or multi-owner. Table 3 shows that single-owner accounts are dominant, accounting for 79% of all accounts. Participants explained how they determined ownership by selecting (multiple selection allowed) from a list of factors derived from the pilots (Figure 5); who paid for the account (635 accounts, 75.3%) contributed to ownership most often, followed by who initially created the account (518 accounts, 61.4%).

We found that in both single-owner and multi-owner shared accounts, creation of shared passwords is primarily an individual process. Only 39 (4.6%) accounts involved two or more people in password creation, whereas individuals created passwords for 747 of the accounts (88.6%). Password generators were used to create passwords for 34 accounts (4.0%). For the remaining 23 accounts (2.7%), participants reported that they were either unsure of how many users were involved in password creation or elected not to answer. Overwhelmingly, account owners are responsible for creating passwords for shared accounts. Across both single-owner and multi-owner accounts, cases in which non-owner users were involved in password creation (either creating the password by themselves or collaboratively with other users) are few, amounting to only 41 accounts overall (4.9%).

While password creation is primarily handled by one person, password-makers often take into account the capabilities of other users, especially in the case of young or elderly users. In these situations, usability may be prioritized over security. For example, Participant 16 described the password to a streaming account as "very basic" because "My grandparents are, well, grandparents. I wanted to make sure they didn't have any more difficulties getting it set up than they needed." In creating a shopping account password, Participant 96 said,
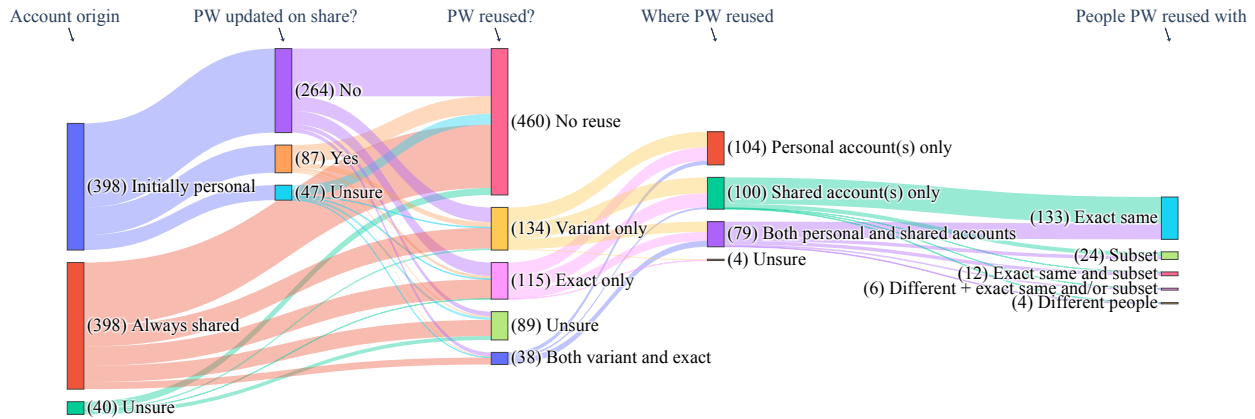
Figure 6: Origins of shared accounts and incidents of password reuse

"security was mildly important, but mostly I wanted an older not-so-computer-savvy relative to be able to enter it correctly." For a shared streaming service, Participant 222 felt that "It is more important for this account to ensure everyone's access than to make the password complex."

4.3 Account origins and password reuse

Shared accounts are created with the intention of sharing as often as not. Users frequently reuse passwords from these shared accounts both in other shared accounts and in personal, non-shared accounts.

Accounts may be created directly for the purpose of sharing or start as a personal (unshared) account; Figure 6 summarizes the sharing and reuse life cycle. We found that 398 (47.2%) of the 843 accounts that participants described in-depth were shared accounts from the start, while the same number, 398 accounts (47.2%), began as personal accounts and were later shared with others. For the remaining accounts, participants were either unsure or elected not to answer.

Account owners often fail to update passwords when personal accounts become shared. Of the 398 shared accounts that began as personal accounts, 264 accounts (66.3%), representing 158 participants (52.7%), did not have their passwords updated when they became shared. Password reuse is also frequent. Nearly half of our participants, (136, 45.3%), reused a shared password elsewhere. In all, 287 shared accounts (34.0%) have their passwords reused in some manner.

Some people reuse passwords verbatim (115 accounts, 13.6%) while others reuse a variant of the password (134 accounts, 15.9%). Because people reuse passwords across multiple accounts, they may also employ a combination of these strategies (38 accounts, 4.5%). Shared passwords see roughly equal reuse in other shared accounts (179 accounts) and in personal accounts (183 accounts).

For passwords reused among multiple shared accounts, the passwords of 133 accounts (15.8%), representing 69 partic-

ipants (23%), are reused in some manner among accounts shared with the exact same people. However, the passwords for 46 accounts (5.4%), representing 27 participants (9%), are reused among accounts shared by different groups of people. Because this may include subsets or supersets of the original group, or even different people entirely, knowledge of the original user's password or creation strategy may spread widely.

Another risk comes from reusing shared passwords (exact or variant-of) on personal accounts, which was reported by 96 participants (32%). In particular, 104 shared accounts (12.3%) have passwords in common only with personal accounts, and 79 shared accounts (9.4%) have passwords in common with both personal and other shared accounts. In total, 183 shared accounts (21.7%) have their passwords directly or indirectly used for personal accounts.

When explaining password reuse, many participants expressed sentiments similar to those observed in single-user accounts in prior work. This includes reuse to cope with the sheer number of passwords they are expected to remember [15] ("I made it similar to other passwords I have because I cannot remember a bunch of passwords to save my life," Participant 4); rationing effort in accounts perceived to be low-value [46] ("I don't want to have to add seven passwords to my list of different passwords I use, especially not for something of low importance," Participant 118); modifying old passwords perceived to be secure as a means of improving recall [48] ("Variation of a password from my college days, it's pretty much stuck in my memory and complex enough for me to feel safe using it," Participant 63); and reusing passwords for thematically similar accounts [44] ("The password is used on a bundle of streaming services, all with the same password," Participant 250).

However, other participants discussed reusing passwords for the purpose of improving usability of multiple accounts shared among a group. Many reasoned that reusing the same password for all accounts they shared with this common group

Distribution Mechanism	Count	Percent
Verbal	489	58.0%
E-mail/text/IM	332	39.4%
Manually entered for recipient	27	3.2%
Written on physical document	24	2.8%
Via password manager	17	2.0%

Table 4: Password distribution mechanisms. Count may add up to more than 843 due to multi-selection. "Prefer not to answer" not shown.

would circumvent the need for each user to individually remember separate passwords for each account. For instance, Participant 95 said, "...it was easy for our family and users to remember it because it had been previously used for another account we all used." Participant 17 wrote that one of their shared account passwords is "...a variation of the password that we use on all the other accounts we share so that it is easy to remember." Finally, Participant 179 described using the "same password for all shared accounts so that the people sharing it can easily log in."

4.4 Other security behaviors for shared accounts

Two-factor authentication is uncommon in shared accounts. Passwords are often transmitted verbally, and forgotten passwords have the potential to disrupt access or cause conflict in the sharing process.

Two-factor authentication. Only about one fifth of shared accounts surveyed, 174 accounts (20.6%), had two-factor authentication (2FA) enabled. Of the 477 accounts participants said did not have 2FA, users actively disabled 2FA for 36 accounts (7.5%) in order to facilitate sharing. For the remaining 192 accounts, participants were unsure if 2FA was enabled.

Password distribution and retrieval. We asked participants how they distributed or received passwords for the accounts they shared (Table 4). Most of the time, they simply read the password out loud. We hypothesize that this may result in simpler or more pronounceable passwords being favored, though this question requires more research.

We also asked participants what they would do in the event that they forgot the password to their shared account (Table 5). Account owners tended to favor resetting the password, while non-owners most favored asking the owner for the password. Either strategy has the potential to disrupt access or cause conflict in the sharing process: password resets can lock other users out of the account if the new password is not re-distributed, and some account owners expressed that they

Retrieval Mechanism	Frequency	
	Owner	Non-owner
Reset password	66.7% (342)	15.1% (50)
Ask another user	22.4% (125)	18.4% (61)
Ask owner/co-owner	12.1% (62)	74.6% (247)
Refer to password distribution message	4.9% (25)	9.7% (32)
Refer to written document	3.7% (19)	1.0% (3)
Refer to password manager	4.7% (24)	1.0% (3)
Guess until gain access	0.0% (0)	<1.0% (1)
Give up access	0.0% (0)	<1.0% (1)

Table 5: What participants would do if they forgot the password to their shared account. Count may add up to more than 843 due to multi-selection.

did not want other users to repeatedly ask them for the password ("I do not want to be bothered each time the password is forgotten," Participant 197).

4.5 Comparison with personal accounts

Participants tend to employ similar password creation strategies for both personal and shared accounts, though account accessibility influences creating passwords for shared accounts. Some, but not all, participants attempt to avoid cross-contamination of personal and shared passwords.

Next, we compare how people treat shared and personal accounts. We focused our analysis on the subset of shared accounts in which participants reported being directly involved in password creation (494 accounts, 232 participants). We compared these to analogous personal accounts and asked each participant for examples of these, if they had any, limiting them to one per account type (as defined in Table 2). Overall, 198 participants described at least one personal account, for a total of 230 personal accounts (Table 6).

Factors important to password creation. For both personal and shared accounts, we asked participants to rate on a five-point Likert-style scale how important the following six factors were for creating their passwords:

- Having a complex password
- Having a memorable (to me) password
- Having a password that is hard to guess
- Having a long password
- Having a password unique from my other passwords
- Being able to store the password in a password manager

We binned these Likert ratings into binary variables for analysis (neutral ratings were grouped with those indicating low importance). To check for correlation between these

Account Type	Count	Percent
Video/Music Streaming	158	68.7%
Shopping	49	21.3%
Finances	9	3.9%
Rent/Utilities	7	3.0%
Health Insurance/Services	2	0.9%
Gaming	2	0.9%
Social Media	1	0.4%
File Sharing	1	0.4%
Other	1	0.4%

Table 6: Types and counts of personal accounts described by participants in "personal accounts" (230 accounts total).

Variable	Odds Ratio	Conf. Int.	*p*-value
Memorable (to me)	1.2	[0.8 - 1.7]	0.350
Hard to Guess	0.6	[0.4 - 0.9]	0.020*
Unique	1.6	[1.1 - 2.3]	0.016*
PW Manager	1.0	[0.7 - 1.4]	0.889

Table 7: Binomial logistic mixed-effects regression on participants' Likert ratings for factors important to password creation in personal and shared accounts. Pseudo-R^2 = 0.02 using the Aldrich-Nelson method [14]. Odds ratios above 1 indicate higher likelihood of the variable being rated as important in a shared account compared to a personal account.

binary factors, we calculated the tetrachoric correlation coefficient, appropriate for correlating binary data, between each factor pair [11]. Three factors connected to password composition and strength ("Having a complex password," "Having a long password," "Having a password that is hard to guess") were all highly correlated ($|r_{tc}| > 0.8$). As participants did not rate the importance of these factors differently, we decided to keep the most general of the three, "Having a password that is hard to guess," and exclude the others ("long" and "complex" passwords) from the analysis.

To identify factors differing between shared and personal accounts, we then constructed a generalized linear mixed-effects model (binomial logistic). The dependent variable was if an account was personal or shared; independent variables included the account type (categorical) and the four remaining creation factors (binary). We compared potential models by testing all possible combinations of covariates and selected our final model based on minimum Akaike Information Criterion (AIC) [2]. Table 7 shows the final model. Odds ratios above 1 correspond to increased importance for shared accounts relative to the baseline (personal accounts).

The final model showed that participants were 1.6× more likely to rate password uniqueness as important for shared accounts than for personal accounts. Conversely, participants

were 1.7× more likely ($\frac{1}{0.6}$) to rate low guessability as important for personal accounts than shared accounts. These results accord with our other findings: participants do not necessarily prioritize security as highly for shared accounts, but are somewhat more concerned about limiting reuse when sharing a password. Memorability and ability to use a password manager did not show a significant difference between personal and shared accounts; account type was dropped from the final model during model selection.

Password creation strategies and motivations. We asked participants to describe their strategies for creating their passwords for personal and shared accounts. Tables 8 and 9 describe the most popular strategies and motivations. Full codebooks are available in the Appendix (Tables 10 and 11).

Behaviors most commonly reported for both personal and shared accounts reflect password behaviors highlighted in previous literature on single-user accounts. These include incorporating meaningful information like birthdays and names of pets [7], reusing passwords with or without slight modifications [10], and attempting to balance security and memorability of passwords [15, 46].

However, there are a few key differences. Some strategies and motivations more frequently discussed for personal accounts included: password generators and managers, relating the password to the service itself, and following personal algorithms for password generation. In contrast, making the password simpler or easier to use was more commonly reported for shared passwords.

Participants' explanations offer insights explaining these differences. While improving password recall and account security served as common motivators in both personal and shared accounts, participants were more often concerned with making the account easy to access in the shared account scenario. This takes several forms, such as making a password that could be easily entered by all users across different devices such as phones, computers, tablets, and even gaming consoles ("I wanted something simple I could give to my wife so she could watch Netflix on her iPad or use at school on occasion," Participant 62; "We needed a password that we could easily enter using the different interfaces where it is used," Participant 223). Other reasons included simplifying passwords for elderly users and children (as highlighted in Section 4.2) or reusing passwords among accounts shared by members of a group, as discussed in Section 4.3. In addition, for 48 of the 216 (22.2%) shared accounts where improving password recall served as a motivator, password-makers specifically stated that they wanted the password to be memorable for all users on the account rather than just themselves.

Threat models for shared accounts. When discussing security in shared accounts, participants more often expressed concerns over external threats to their accounts, such as hackers ("It's a utility and could be targeted by hackers," Partic-

Code	Frequency		Definition
	Shared	**Personal**	
Meaningful info.	22.5% (111)	23.0% (53)	Used information that is meaningful to at least one user
Memorable	20.6% (102)	14.8% (34)	Prioritized making the password memorable
Reuse	16.8% (83)	20.0% (46)	Reused (either exactly or a variant of) another password
Secure	14.0% (69)	12.6% (29)	Prioritized making the password secure
Personal algorithm	13.8% (68)	18.7% (43)	Used a personal algorithm for passwords, such as minimum rules or a pattern of units (e.g., numbers-word-numbers)
Password generator	8.5% (42)	10.4% (24)	Used a password generator (i.e., one in a password manager)
Random	5.1% (25)	4.8% (11)	Created randomly without the use of a password generator
Related to service	1.4% (7)	6.1% (14)	Related to the service that the account is for

Table 8: Common (used in more than 5% of either shared or personal accounts surveyed) password generation strategies used by participants for accounts where they were involved in password creation. Participants sometimes indicated more than one strategy per password.

Code	Frequency		Definition
	Shared	**Personal**	
Recall	43.7% (216)	53.5% (123)	Wanted the password to be easy to recall for users
Secure account	35.4% (175)	43.0% (99)	Prioritized the security of the account
Easy to access account	8.5% (42)	2.6% (6)	Wanted the account to be accessed easily
Easy to make	5.9% (29)	4.3% (10)	Password was easy to make

Table 9: Common (used in more than 5% of either shared or personal accounts surveyed) motivations for participants' choice of password strategy. Participants sometimes indicated more than one motivation per password.

ipant 61; "Amazon is a target for thieves, so I want to be as careful as possible," Participant 96) rather than internal threats. Nonetheless, a few users took precautions to avoid cross-contaminating passwords between shared and personal accounts, even if they did not directly refer to other users as potential security threats ("Because I am sharing this password and don't want it to be the same password I use for other things," Participant 48; "It keeps my other accounts secure as it is unique to this account," Participant 135).

Effort rationing based on perceived account value. Similar to previous work on effort rationing [44], participants discussed conserving effort for accounts they deemed as more sensitive ("I wanted this password to be harder to guess because it's attached to my bank accounts," Participant 142; "More complicated password since a shopping site," Participant 156) and deferring to weaker security practices for "less valuable" accounts ("There isn't much that a hacker could do in this account, so security is not as important for this [travel account]," Participant 233; "Minimal loss if password gets stolen, will just reset, and no real way it can cost us money/security," Participant 203). Sensitive accounts often included shopping accounts, which can have stored credit cards, and utilities accounts, which are associated with physical res-

idences. On the other hand, streaming accounts were often considered to be less valuable by participants due to storing limited information or having little impact if compromised.

5 Discussion

In contrast to prior works, which examine password creation in single-user accounts or account sharing behavior post-password creation, we combine the two and study creation and use of shared passwords in multi-user accounts. Our study sheds light on how people share passwords and offers important implications for system developers.

5.1 Comparisons with the single-user context

Similarities. We observe that usability challenges highlighted by prior works in single-user contexts influence creation of shared passwords in similar ways. As discussed in Sections 4.3 and 4.4, people often reuse passwords in shared accounts to cope with having more passwords than they (or the people they are sharing with) can effectively remember, ration effort spent on low-value accounts, and use old, "secure" passwords as a resource for deriving new passwords [15, 44, 46, 48]. Our participants engaged in common

behaviors previously seen in the single-user setting for both passwords they intend to share and those they do not, such as using meaningful data in their passwords [7] and relying on personal algorithms to create passwords [46].

Differences. We observe that elements specific to the context of sharing exert unique pressures on password-makers. The presence of multiple users may exacerbate the security-usability trade-off; users discuss making deliberate security concessions when creating passwords with the intention of sharing in order to proactively improve usability, such as by making simpler and easier to remember passwords when sharing with young or old users. Further, account sharing engenders a particular type of password reuse: a common password among multiple accounts shared by a group of users.

5.2 Shared passwords created by a single user

Password-making in shared accounts is an individual process, often performed solely by an account owner, rather than as a collaborative effort between users. We note that a similar dynamic has been observed in the context of smart homes, where users who install home smart devices have an outsized role in controlling configurations and repair of these devices [13]. This individual effort comes in two varieties, each with their own unique security implications.

As discussed above, when users create passwords *with* the intention of sharing, they may make security concessions for the sake of usability. When users create passwords *without* the intention of sharing, as in the case of personal accounts that did not have their passwords updated when they were shared, the password inherits the typical mechanisms and weaknesses of other personal account passwords, including the potential for password reuse. In about two thirds of reused passwords (63.7%), participants stated they reused these passwords in both shared and personal accounts. This form of reuse can create additional vulnerabilities for personal passwords when the account becomes shared, increasing the opportunity for the password to be phished, leaked, or otherwise stolen and then used in credential-stuffing attacks.

Because account owners play the primary role in password creation and shared account management, they represent a promising target to improve safety in sharing.

5.3 Implications for developers

Most systems are still designed with the assumption of one person per account. While services may quietly tolerate or outright prohibit account sharing [28, 31], it is nonetheless a common behavior often undertaken for reasons that are important to the user, and as such it will likely continue despite its potential security risks. Technology designers and systems developers should keep this reality in mind and tailor systems and advice to minimize potential harms from sharing.

Account sharing without credential sharing. Services that tolerate or encourage sharing can enable account sharing without password sharing, for example by giving each user of a shared account unique credentials to reduce unintentional propagation of personal passwords. However, the overhead of creating such sub-accounts and associated passwords, as well as untangling and migrating individual user data if this schema were to be applied to existing accounts being shared, may deter users, and they may instead choose to default to using a single shared account, as seen in the smart home setting [23]. Researchers should instead investigate more usable access control alternatives for account sharing.

Helping users discern who is accessing a shared account. Some participants reported they were unsure exactly how many people had access to some of their accounts. It would be helpful for developers to provide a simple and comprehensible view of login history: when and where an account has been used from, together with the ability to annotate logins and associate them with specific users. These account security indicators can alert users to unwanted access and help owners remove no-longer-authorized users [9]. We also find that people tend to reuse group passwords within (approximately) the same group of people; a login history might help users to understand whether a group password has been compromised or has traveled beyond its initial intended recipients.

Password managers. Not all services want to encourage account sharing, including for financial or security reasons. Password managers could play a greater role in enabling sharing while maintaining security and reducing password reuse on services that do not wish to implement account sharing features themselves. Several participants reported using a password manager to distribute and store shared passwords. Indeed, some popular password managers offer family plans and one-time password sharing options that claim to simplify sharing and security [1, 22, 29]. In addition, password managers enable non-owner users to retrieve forgotten passwords without inconveniencing the account owner, synchronize passwords and account access across devices without having to reenter passwords, and generate passwords that are easier to verbalize yet retain security. A number of participants reported disabling 2FA in order to facilitate account sharing; password managers can enable sharing of 2FA-protected accounts through sharing time-based one-time password seeds. However, the usability of these password managers in the context of account sharing has yet to be evaluated. Usability challenges already represent a major hurdle to adoption of password managers in the single-user context [35], and these challenges may be further exacerbated in the shared account setting where users prioritize account accessibility even more.

Trust and social norms. Our participants' account sharing behavior highlights the importance of trust and social norms in maintaining account security while engaging in insecure behavior. Participants primarily shared accounts with trusted family members and partners. While some took precautions to protect themselves from other users, such as by avoiding cross-contamination between shared and personal passwords, many others (almost a third of our participants) did not. Participants more often cited external threats (hackers) over internal threats (other users) when discussing security. While our work centers on users in the U.S., research in other cultural contexts have similarly highlighted the role of trust and social norms in maintaining account security while sharing passwords [3, 4].

This reliance on trust for security can serve as a double-edged sword: users may reuse passwords or disable security measures like 2FA. In the case of reusing passwords from personal accounts on shared accounts, users may be granting others access to accounts they do not intend to share on a technical level, but trust them not to access these accounts or engage in harmful behavior. Similar behaviors have also been observed in the context of smart homes; device owners often rely on trust with other users and social norms rather than strict access controls for security [23, 53]. Interventions will need to find a way to maintain account security if this trust can no longer be relied upon, as in the case of relationships ending. On the other hand, previous literature has suggested that these relationships and group dynamics can be leveraged to improve security behavior of members less versed in secure behaviors [51]. In the account sharing scenario, groups can perhaps encourage adoption of other secure habits among their members, such as the use of password managers or other secure password behaviors.

5.4 Future work

Sharing of highly sensitive accounts. There remain some important open questions about shared accounts and passwords. Our study surveyed sharing of accounts broadly; however, some accounts have greater security implications, such as those for financial institutions and utility companies. While our participants more often reported employing secure strategies like randomization of passwords (versus using meaningful information) for shared financial accounts compared to less sensitive accounts like streaming accounts, we lacked a sufficiently large sample to draw definitive conclusions. Due to their importance, future work could investigate these sensitive accounts, their password strategies, and relative security of these shared passwords specifically.

Password reuse in accounts of varying sensitivity. Our study uncovered a high degree of password reuse between shared and personal accounts. While this is concerning, more information is needed to fully understand the security implications of this reuse. Do the reused passwords span both low-

and high-value accounts? Do these personal accounts have additional protection measures (e.g., two-factor authentication), or are they accessible by anyone with the password? Do people understand the ramifications of their password-sharing choices? Future researchers could investigate these questions as well as others about the mental models of those engaging in reuse of shared passwords.

Verbalization of shared passwords. We found that 58.0% of shared passwords in this study were transmitted verbally. We hypothesize that one reason people create weaker passwords for shared accounts is to make them easy to communicate. Future work could test this supposition directly by investigating the relationship between the distribution mechanism and password composition, as well as test the usability of generators that claim to make secure, verbalizable passwords.

Ending account sharing. Previous literature has highlighted that ending sharing and updating passwords to all formerly-shared accounts is a tedious and challenging process for users [30, 34], and we posit that the cross-contamination of passwords between shared and personal accounts discussed by our participants would further amplify the difficulties users face when attempting to end account sharing.

Other sharing contexts. Our participant pool primarily shared accounts with family members and friends. Future work could examine shared password creation in other contexts that have different security implications and different trust dynamics, such as in the workplace.

6 Conclusion

We conducted a U.S. census-representative survey ($n = 300$) to understand password creation in shared accounts. We found that the typical user tends to share accounts with partners and family members, and streaming accounts are most commonly shared. Creation of shared passwords is predominantly an individual rather than collaborative process, typically performed by the account owner. While users mostly employ similar strategies to create both shared and non-shared passwords, prioritization of usability of shared accounts can lead to deliberate security concessions. Password reuse is common, occurring in roughly a third of shared accounts surveyed. Accounts shared by a group are often accessible by a single common password. Among shared accounts with reused passwords, approximately two-thirds of these passwords, representing a third of our participants, are reused in some manner on personal accounts. Technology creators and security experts should take these findings—and the inevitable reality of credential sharing—into account and design systems that can support sharing while minimizing harm.

7   Acknowledgments

References

[1] 1Password. 1password families, Accessed 2023-05-03. URL https://1password.com/affiliate/families.

[2] Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected papers of Hirotugu Akaike*. Springer, 1998.

[3] Aniqa Alam, Elizabeth Stobert, and Robert Biddle. "this is different from the western world": Understanding password sharing among young bangladeshis. In *Symposium on Usable Security and Privacy (USEC)*, 2023.

[4] Deena Alghamdi, Ivan Flechais, and Marina Jirotka. Security practices for households bank customers in the kingdom of saudi arabia. In *Symposium On Usable Privacy and Security (SOUPS)*, 2015.

[5] Jennifer L. Bevan. Social networking site password sharing and account monitoring as online surveillance. In *Cyberpsychology, Behavior, and Social Networking*, volume 21, pages 797–802, 2018.

[6] Joseph Bonneau, Cormac Herley, Paul C. van Oorschot, and Frank Stajano. The quest to replace passwords: A framework for comparative evaluation of web authentication schemes. In *IEEE Symposium on Security and Privacy (SP)*, 2012.

[7] Kay Bryant and John Campbell. User behaviours associated with password security and management. In *Australasian Journal of Information Systems*, volume 14, 2006.

[8] Karoline Busse, Julia Schäfer, and Matthew Smith. Replication: No one can hack my mind revisiting a study on expert and Non-Expert security practices and advice. In *Symposium on Usable Privacy and Security (SOUPS)*, 2019.

[9] Alaa Daffalla, Marina Bohuk, Nicola Dell, Rosanna Bellini, and Thomas Ristenpart. Account security interfaces: Important, unintuitive, and untrustworthy. In *USENIX Security Symposium (USENIX Security)*, 2023.

[10] Anupam Das, Joseph Bonneau, Matthew Caesar, Nikita Borisov, and Xiaofeng Wang. The tangled web of password reuse. In *Network and Distributed System Security (NDSS) Symposium*, 2014.

[11] D. R. Divgi. Calculation of the tetrachoric correlation coefficient. 44(2):169–172, 1979.

[12] Serge Egelman, A.J. Bernheim Brush, and Kori M. Inkpen. Family accounts: A new paradigm for user accounts within the home environment. In *Conference on Computer Supported Cooperative Work (CSCW)*, 2008.

[13] Christine Geeng and Franziska Roesner. Who's in control? interactions in multi-user smart homes. In *Conference on Human Factors in Computing Systems (CHI)*, 2019.

[14] Timothy M Hagle and Glenn E Mitchell. Goodness-of-fit measures for probit and logit. *American Journal of Political Science*, 1992.

[15] Ameya Hanamsagar, Simon S. Woo, Chris Kanich, and Jelena Mirkovic. Leveraging semantic transformation to investigate password habits and their causes. In *Conference on Human Factors in Computing Systems (CHI)*, 2018.

[16] Philip G. Inglesant and M. Angela Sasse. The true cost of unusable password policies: Password use in the wild. In *Conference on Human Factors in Computing Systems (CHI)*, 2010.

[17] Iulia Ion, Rob Reeder, and Sunny Consolvo. "...no one can hack my mind": Comparing expert and non-expert security practices. In *Symposium On Usable Privacy and Security (SOUPS)*, 2015.

[18] Jyun-Yu Jiang, Cheng-Te Li, Yian Chen, and Wei Wang. Identifying users behind shared accounts in online streaming services. In *Conference on Research & Development in Information Retrieval*, 2018.

[19] Joseph 'Jofish' Kaye. Self-reported password sharing strategies. In *Conference on Human Factors in Computing Systems (CHI)*, 2011.

[20] Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology*. Fourth edition edition, 2023.

[21] Airi M I Lampinen. Account sharing in the context of networked hospitality exchange. In *Conference on Computer Supported Cooperative Work (CSCW)*, 2014.

[22] LastPass. Lastpass families, Accessed 2023-05-03. URL https://www.lastpass.com/products/family-password-manager.

[23] Nathan Malkin, Alan F. Luo, Julio Poveda, and Michelle L. Mazurek. Optimistic access control for the smart home. In *IEEE Symposium on Security and Privacy (SP)*, 2023.

[24] Tara Matthews, Kerwell Liao, Anna Turner, Marianne Berkovich, Robert Reeder, and Sunny Consolvo. "she'll just grab any device that's closer": A study of everyday device & account sharing in households. In *Conference on Human Factors in Computing Systems (CHI)*, 2016.

[25] Mary L. McHugh. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*, 22(3), 2012.

[26] Helena M. Mentis, Galina Madjaroff, and Aaron K. Massey. Upside and downside risk in online security for older adults with mild cognitive impairment. In *Conference on Human Factors in Computing Systems (CHI)*, 2019.

[27] Collins W. Munyendo, Yasemin Acar, and Adam J. Aviv. "in eighty percent of the cases, i select the password for them": Security and privacy challenges, advice, and opportunities at cybercafes in kenya. In *IEEE Symposium on Security and Privacy (SP)*, 2023.

[28] Netflix. An update on sharing, 2023-02-08. URL https://about.netflix.com/en/news/an-update-on-sharing.

[29] NordPass. Boost your business security with ease, Accessed 2023-05-03. URL https://nordpass.com/nordpass-business-solution.

[30] Borke Obada-Obieh, Yue Huang, and Konstantin Beznosov. The burden of ending online account sharing. In *Conference on Human Factors in Computing Systems (CHI)*, 2020.

[31] Kate O'Flaherty. The disney password sharing crackdown is about to begin, 2023-08-11. URL https://www.forbes.com/sites/kateoflahertyuk/2023/08/11/the-disney-password-sharing-crackdown-is-about-to-begin/?sh=3e14dae2577c.

[32] Kenneth Olmstead and Aaron Smith. Password management and mobile security. 2017.

[33] Joris Van Ouytsel. The prevalence and motivations for password sharing practices and intrusive behaviors among early adolescents' best friendships – a mixed-methods study. In *Telematics and Informatics*, volume 63, page 101668, 2021.

[34] Cheul Young Park, Cori Faklaris, Siyan Zhao, Alex Sciuto, Laura Dabbish, and Jason Hong. Share and share alike? an exploration of secure behaviors in romantic relationships. In *Symposium on Usable Privacy and Security (SOUPS)*, 2018.

[35] Sarah Pearman, Shikun Aerin Zhang, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. Why people (don't) use password managers effectively. In *Symposium on Usable Privacy and Security (SOUPS)*, 2019.

[36] Hirak Ray, Flynn Wolf, Ravi Kuber, and Adam J. Aviv. Why older adults (don't) use password managers. In *USENIX Security Symposium (USENIX Security)*, 2021.

[37] Elissa M. Redmiles, Sean Kross, and Michelle L. Mazurek. How well do my results generalize? comparing security and privacy survey results from mturk, web, and telephone samples. In *IEEE Symposium on Security and Privacy (SP)*, 2019.

[38] Johnny Saldaña. *The coding manual for qualitative researchers.* Sage Publications Ltd, 2009.

[39] Nithya Sambasivan, Garen Checkley, Amna Batool, Nova Ahmed, David Nemer, Laura Sanely Gaytán-Lugo, Tara Matthews, Sunny Consolvo, and Elizabeth Churchill. "privacy is not for me, it's for those rich women": Performative privacy practices on mobile phones by women in south asia. In *Symposium on Usable Privacy and Security (SOUPS)*, 2018.

[40] Richard Shay, Saranga Komanduri, Patrick Gage Kelley, Pedro Giovanni Leon, Michelle L. Mazurek, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. Encountering stronger password requirements: user attitudes and behaviors. In *Symposium on Usable Privacy and Security (SOUPS)*, 2010.

[41] Richard Shay, Saranga Komanduri, Patrick Gage Kelley, Pedro Giovanni Leon, Michelle L. Mazurek, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. Encountering stronger password requirements: User attitudes and behaviors. In *Symposium on Usable Privacy and Security (SOUPS)*, 2010.

[42] Lucy Simko, Ada Lerner, Samia Ibtasam, Franziska Roesner, and Tadayoshi Kohno. Computer security and privacy for refugees in the united states. In *IEEE Symposium on Security and Privacy (SP)*, 2018.

[43] Yunpeng Song, Cori Faklaris, Zhongmin Cai, Jason I. Hong, and Laura Dabbish. Normal and easy: Account sharing practices in the workplace. 2019.

[44] Elizabeth Stobert and Robert Biddle. The password life cycle: User behaviour in managing passwords. In *Symposium on Usable Privacy and Security (SOUPS)*, 2014.

[45] Jenny Tang, Eleanor Birrell, and Ada Lerner. Replication: How well do my results generalize now? the external validity of online privacy and security surveys. In *Symposium on Usable Privacy and Security (SOUPS)*, 2022.

[46] Blase Ur, Fumiko Noma, Jonathan Bees, Sean M. Segreti, Richard Shay, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. "i added '!' at the end to make it secure": Observing password creation in the lab. In *Symposium on Usable Privacy and Security (SOUPS)*, 2015.

[47] Blase Ur, Jonathan Bees, Sean M. Segreti, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. Do users' perceptions of password security match reality? In *Conference on Human Factors in Computing Systems (CHI)*, 2016.

[48] Emanuel von Zezschwitz, Alexander De Luca, and Heinrich Hussmann. Survival of the shortest: A retrospective analysis of influencing factors on password composition. In Paula Kotzé, Gary Marsden, Gitte Lindgaard, Janet Wesson, and Marco Winckler, editors, *Human-Computer Interaction – INTERACT 2013*, pages 460–467. Springer Berlin Heidelberg, 2013.

[49] Noel Warford, Tara Matthews, Kaitlyn Yang, Omer Akgul, Sunny Consolvo, Patrick Gage Kelley, Nathan Malkin, Michelle L. Mazurek, Manya Sleeper, and Kurt Thomas. Sok: A framework for unifying at-risk user research. In *IEEE Symposium on Security and Privacy (SP)*, 2022.

[50] Rick Wash, Emilee Rader, Ruthie Berman, and Zac Wellmer. Understanding password choices: How frequently entered passwords are re-used across websites. In *Symposium on Usable Privacy and Security (SOUPS)*, 2016.

[51] Hue Watson, Eyitemi Moju-Igbene, Akanksha Kumari, and Sauvik Das. "we hold each other accountable": Unpacking how social groups approach cybersecurity and privacy together. In *Conference on Human Factors in Computing Systems (CHI)*, 2020.

[52] Monica Whitty, James Doodson, Sadie Creese, and Duncan Hodges. Individual differences in cyber security behaviors: an examination of who is sharing passwords. In *Cyberpsychology, behavior and social networking*, volume 18, pages 3–7, 2015.

[53] Eric Zeng and Franziska Roesner. Understanding and improving security and privacy in Multi-User smart homes: A design exploration and In-Home user study. In *USENIX Security Symposium (USENIX Security)*, 2019.

[54] Yinqian Zhang, Fabian Monrose, and Michael K. Reiter. The security of modern password expiration: An algorithmic framework and empirical analysis. In *Conference on Computer and Communications Security (CCS)*, 2010.

A Supplementary Figures
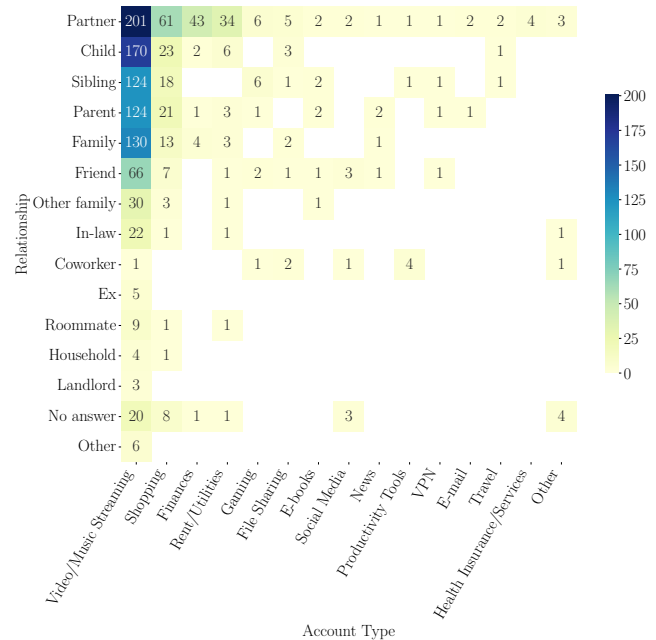
Figure 7: Who participants report sharing their accounts with in "Shared accounts overview," separated by account type.

B Survey Instrument

B.1 Shared accounts overview

In this survey, we will be asking you about the accounts you share with other people.

By **account**, we mean any website or system where you log in with a username (or email address) and password combination in order to access services or content.

People often share accounts (such as those for streaming, shopping, and finances) for a wide variety of reasons.

By **shared account**, we mean any account where you and at least one other person both use the same username (or email address) and password combination in order to access and use the account, either at the same time or taking turns.

This EXCLUDES: Accounts where each person uses a different username (or email) and password combination to log in. Accounts where only a username or password is needed (but not both), such as shared Wi-Fi.

1. How many accounts do you currently share with at least one other person and **can currently log into or access?**

   Using one account per line, please describe them below. If you have multiple accounts for one type of service (such as multiple streaming accounts) that you share with others, please describe them separately.

	What website or service is this account for? Feel free to describe the type of service if you would rather not name the account.	What option best describes this account?	Excluding yourself, how many people do you share this account with (to the best of your knowledge)?	Check this column if you are NOT certain how many people share this account.	Who do you share this account with? For example: friends, siblings, coworkers, etc.
Account 1	e.g. "Netflix"			☐	e.g. "parents"
Account 2				☐	
Account 3				☐	
Account 4				☐	
Account 5				☐	
Account 6				☐	
Account 7				☐	
Account 8				☐	
Account 9				☐	
Account 10				☐	

2. If you have any other accounts that you share with other people, please describe each account here with:

    (a) The name of the website/service the account is for.

    (b) How many people, excluding yourself, use this account.

    (c) Who you share this account with.

B.2 Personal accounts

This section is repeated once for every [account type] that the participant reports having

a shared account of in Part A

3. Think about the accounts you **DO NOT** share with anyone else (you are the only person with access to these accounts and the only one that knows the username/password combination).

   Do you have any **[account type]** accounts that you **DO NOT** share with anyone else?
   - ○ Yes - I have a [account type] account that I DO NOT share with anyone else
   - ○ No - I do not have such a [account type] account
   - ○ Not sure / Prefer not to answer

The remaining questions in this section are only shown if the participant answers "Yes"

to the above question

4. Think about one such [account type] account that you DO NOT share with anyone else. What website or service is this account for?

5. Think about the password you made for this account. How important were each of the following factors in creating your password?

	1 - Not important at all	2	3	4	5 - Extremely important	Prefer not to answer
Having a complex password	○	○	○	○	○	○
Having a memorable (to me) password	○	○	○	○	○	○
Having a password that is hard to guess	○	○	○	○	○	○
Having a long password	○	○	○	○	○	○
Having a password that is unique from my other passwords	○	○	○	○	○	○
Being able to store the password in a password manager	○	○	○	○	○	○
Other	○	○	○	○	○	○

6. Think about the password you made for this account. Did you use any of the following strategies to create your password? Check all that apply.

   If you are unsure of the password or don't remember it exactly, please check the "I do not remember the exact password to this account" option.
   - ☐ Based on the name of someone or something
   - ☐ Based on a word or name with numbers/symbols added to beginning or end
   - ☐ Based on a word or name with numbers or symbols replacing some letter (e.g. '@' for 'a')
   - ☐ Based on a non-English word
   - ☐ Based on a date
   - ☐ Incorporates a passphrase (e.g. 'correcthorsebatterystaple')
   - ☐ Based on meaningful information to you (e.g. names, favorite things, inside jokes)
   - ☐ Uses lowercase letters
   - ☐ Uses uppercase letters
   - ☐ Longer than 8 characters
   - ☐ Reused a password I use elsewhere
   - ☐ Modified a password I use elsewhere
   - ☐ Follows a password pattern I use elsewhere
   - ☐ Created with a password manager
   - ☐ Intentionally planned to use reset password feature
   - ☐ Easy to read/say
   - ☐ Uses numbers
   - ☐ Uses symbols
   - ☐ Other (please specify) ⎯⎯⎯⎯
   - ☐ I do not remember the exact password to this account
   - ☐ I would prefer not to answer

7. Please briefly explain your overall strategy for making this specific password.
   Please DO NOT tell us your actual password! We are only interested in the strategies you used to come up with your password.

8. Please briefly explain **why** you chose this strategy.

B.3 Shared accounts details

This section is repeated once each of the first four shared accounts that the participant reports in Part A

We'll be asking some questions about the "[account description]" [account type] account that you share with [number shared with] other people, including "[relationship]".

9. Who do you consider to be the owner(s) of this account? Check all that apply.
   - ☐ Myself
   - ☐ One other user
   - ☐ Multiple other users
   - ☐ Other (please specify) ⎯⎯⎯⎯
   - ☐ Not sure / Prefer not to answer

10. How did you determine the owner(s) for this account? Please select all that apply.
    - ☐ Pays for this account
    - ☐ Collects payment from other users
    - ☐ Originally created this account
    - ☐ Uses this account the most
    - ☐ Has access to the email address associated with this account
    - ☐ Other (please specify) ⎯⎯⎯⎯
    - ☐ Not sure / Prefer not to answer

11. Did one person come up with the password for this account, or was it a collaborative effort?
    - ○ One person came up with the password for this account
    - ○ The password for this account was a collaborative effort
    - ○ A password manager or other tool was used to generate the password

○ Other (please describe) ⎯⎯⎯⎯

○ Not sure / Prefer not to answer

12. Who was involved with creating the password to this "[account description]" account? Please select all that apply.

☐ Myself

☐ Other user(s) that I consider to be the owner or joint owner(s)

☐ Other user(s) that I do NOT consider to be the owner or joint owner(s)

☐ Other (please describe) ⎯⎯⎯⎯

☐ Not sure / Prefer not to answer

The following question is only shown if the participant reported that they were involved with password creation (selected "Myself" in the previous question).

13. Think about the password you created or helped create for this "[account description]" account. How important were each of the following factors to you in creating this password?

	1 - Not important at all	2	3	4	5 - Extremely important	Prefer not to answer
Having a complex password	○	○	○	○	○	○
Having a memorable (to me) password	○	○	○	○	○	○
Having a memorable (to the people sharing with me) password	○	○	○	○	○	○
Having a password that is hard to guess	○	○	○	○	○	○
Having a long password	○	○	○	○	○	○
Having a password that is unique from my other passwords	○	○	○	○	○	○
Being able to store the password in a password manager	○	○	○	○	○	○
Other	○	○	○	○	○	○

14. To help us monitor the quality of our data, please select "Somewhat disagree".

○ Strongly disagree

○ Somewhat disagree

○ Neither disagree nor agree

○ Somewhat agree

○ Strongly agree

The following question is only shown if the participant reported that they were involved with password creation.

15. Think about the password you made for [account description]. Did you use any of the following strategies to create your password? Check all that apply.

If you are unsure of the password or don't remember it exactly, please check the "I do not remember the exact password to this account" option.

☐ Based on the name of someone or something

☐ Based on a word or name with numbers/symbols added to beginning or end

☐ Based on a word or name with numbers or symbols replacing some letter (e.g. '@' for 'a')

☐ Based on a non-English word

☐ Based on a date

☐ Incorporates a passphrase (e.g. 'correcthorsebatterystaple')

☐ Based on meaningful information to you (e.g. names, favorite things, inside jokes)

☐ Uses lowercase letters

☐ Uses uppercase letters

☐ Longer than 8 characters

☐ Reused a password I use elsewhere

☐ Modified a password I use elsewhere

☐ Follows a password pattern I use elsewhere

☐ Created with a password manager

☐ Intentionally planned to use reset password feature

☐ Easy to read/say

☐ Uses numbers

☐ Uses symbols

☐ Other (please specify) ⎯⎯⎯⎯

☐ I do not remember the exact password to this account

☐ I would prefer not to answer

The following question is only shown if the participant reported that they were involved with password creation.

16. Please briefly explain your overall strategy for making this specific password. **DO NOT tell us your actual password!** We are only interested in the strategy you used to come up with your password.

The following question is only shown if the participant reported that they were involved with password creation.

17. **Why** did you choose this strategy?.

The following question is only shown if the participant reported that they WERE NOT involved with password creation.

18. Think about the password for [account description]. Does it use any of the following strategies? Check all that apply.

If you are unsure of the password or don't remember it exactly, please check the "I do not remember the exact password to this account" option.

☐ Based on the name of someone or something

☐ Based on a word or name with numbers/symbols added to beginning or end

☐ Based on a word or name with numbers or symbols replacing some letter (e.g. '@' for 'a')

☐ Based on a non-English word

☐ Based on a date

☐ Incorporates a passphrase (e.g. 'correcthorsebatterystaple')

☐ Based on meaningful information to you (e.g. names, favorite things, inside jokes)

☐ Uses lowercase letters

☐ Uses uppercase letters

☐ Longer than 8 characters

☐ Reused a password I use elsewhere

☐ Modified a password I use elsewhere

☐ Follows a password pattern I use elsewhere

☐ Created with a password manager

☐ Intentionally planned to use reset password feature

☐ Easy to read/say

☐ Uses numbers

☐ Uses symbols

☐ Other (please specify) ⎯⎯⎯⎯

☐ I do not remember the exact password to this account

☐ I would prefer not to answer

The following question is only shown if the participant reported that they WERE NOT involved with password creation.

19. What do you think the person(s) creating the password were prioritizing by choosing these strategies? If you are unsure, please give us your best guess.

20. Do you use the password for [account description] on other accounts? Please select all that apply.

☐ Yes - I reuse the password exactly

☐ Yes - I use a variant of this password

□ No - I do not reuse this password anywhere in any form

□ Not sure

□ Prefer not to answer

The following question is only shown if the participant reported that they reused this password (selected "Yes - I reuse the password exactly" and/or "Yes - I use a variant of this password" in the previous question).

21. Do you use the password for [account description] on other accounts? Please select all that apply.
    □ Personal account(s) (not shared with anyone else)

    □ Other shared account(s)

    □ Prefer not to answer

The following question is only shown if the participant reported that they reused this password with other shared accounts (selected "Other shared account(s)" in the previous question).

22. How would you describe the other shared account(s) that use the same password as this account? Check all that apply.
    □ I share the other account(s) with **exactly the same people** as I share this account with

    □ I share the other account(s) with **some, but not all** of the people I share this account with

    □ I share the other account(s) with **people I DO NOT share** this account with

    □ Prefer not to answer

23. Was this account always shared, or did it start as a personal account that later became a shared account?
    ○ This account was always shared

    ○ This account started as an individual account that later became shared

    ○ Not sure

    ○ Prefer not to answer

The following question is only shown if the participant reported that this account began as a personal account (selected "This account started as an individual account that later became shared" in the previous question).

24. When this account became a shared account, was the password changed or updated?
    ○ Yes - the password was updated

    ○ No - the password was not updated

    ○ Not sure

    ○ Prefer not to answer

25. Is two-factor authentication (2FA) currently enabled on this account?
    ○ Yes - two-factor authentication (2FA) is currently enabled on this account

    ○ No - two-factor authentication (2FA) is NOT currently enabled on this account

    ○ Not sure

    ○ Prefer not to answer

The following question is only shown if the participant reported that this account does not have 2FA enabled (selected "No - two-factor authentication (2FA) is NOT currently enabled on this account" in the previous question).

26. Was two-factor authentication (2FA) disabled in order to share this account?
    ○ Yes - two-factor (2FA) authentication was disabled

    ○ No - two-factor (2FA) authentication was NOT disabled

    ○ Not sure

    ○ Prefer not to answer

27. How did you distribute the password to other people sharing this account **OR** how did you receive the password to this account? Please check all that apply.
    □ Verbally (either in-person or over a phone call)

    □ Through e-mail, text, or instant messaging

    □ Other (please specify) _____

    □ Not sure

□ Prefer not to answer

28. If you forgot the password for this account, what would you do? Please check all that apply.
    □ I would use the account's password reset mechanism.

    □ I would ask the account owner for the password.

    □ I would ask another person sharing the account (not the account owner) for the password.

    □ I sent the password to someone else/was originally sent the password through a text message or email, so I would check that.

    □ Other (please specify) _____

    □ Prefer not to answer

B.4 Demographics

29. What is your age? Please type "0" if you prefer not to say.

30. Please select the option that best describes your gender.
    ○ Male

    ○ Female

    ○ Nonbinary

    ○ Another gender (please specify) _____

    ○ Prefer not to say

31. What is your annual household income?
    ○ Less than $25,000

    ○ $25,000 to $34,999

    ○ $35,000 to $49,999

    ○ $50,000 to $74,999

    ○ $75,000 to $99,999

    ○ $100,000 to $149,999

    ○ $150,000 to $199,999

    ○ $200,000 or more

    ○ Prefer not to say

32. Please choose the highest level of education you have completed.
    ○ Have not completed high school

    ○ High school degree or equivalent

    ○ Associate's degree

    ○ Bachelor's degree

    ○ Master's degree

    ○ Professional degree beyond a bachelor's degree (e.g. MD, DDS)

    ○ Doctoral degree

    ○ Prefer not to say

33. Do you have a computer science background? This means working in or holding a degree in computer science or information technology.
    ○ Yes

    ○ No

    ○ Not sure

    ○ Prefer not to say

34. Do you have a background in computer or information security?
    ○ Yes

    ○ No

    ○ Not sure

    ○ Prefer not to say

35. Is there any feedback on our survey or additional information you'd like to provide to help us understand your responses or improve the survey?

C Full Codebooks

| | Frequency | | |
Code	Shared	Personal	Definition
Meaningful info.	22.5% (111)	23.0% (53)	Uses information that is meaningful to at least one user
Memorable	20.6% (102)	14.8% (34)	Prioritized making the password memorable
Reuse	16.8% (83)	20.0% (46)	Reused (either exactly or a variant of) another password
Secure password	14.0% (69)	12.6% (29)	Prioritized making the password secure
Personal algorithm	13.8% (68)	18.7% (43)	Personal algorithm for passwords, such as minimum rules or a pattern of units (e.g., numbers-word-numbers)
Password generator	8.5% (42)	10.4% (24)	Used a password generator (i.e., one in a password manager)
Random	5.1% (25)	4.8% (11)	Created randomly without the use of a password generator
Passphrase	4.7% (23)	4.8% (11)	Passphrase that does not contain any meaningful information
Easy to use	3.2% (16)	0.9% (2)	Easy to use and enter
Unique	3.0% (15)	2.2% (5)	Intentionally avoided reusing an old password or making a similar password
Storage in manager	2.0% (10)	4.3% (10)	Being able to easily store and retrieve their password in a password manager
Related to service	1.4% (7)	6.1% (14)	Related to the service that the account is for
Simple	1.2% (6)	0.0% (0)	Prioritized simplicity
Just meet requirements	1.2% (6)	2.6% (6)	Minimally satisfies the account's password policy
Hard to use	0.8% (4)	0.0% (0)	Cumbersome to use
Another language	0.6% (3)	1.7% (4)	Derived from a non-English language
Environment	0.0% (0)	1.3% (3)	Participant's surroundings were used for parts of the password

Table 10: Password generation strategies used by participants for accounts where they were involved in password creation (494 shared, 230 personal). Participants sometimes indicated more than one strategy per password.

| | Frequency | | |
Code	Shared	Personal	Definition
Recall	43.7% (216)	53.5% (123)	Wanted the password to be easy to recall for users
Secure account	35.4% (175)	43.0% (99)	Prioritized the security of the account
Easy to access account	8.5% (42)	2.6% (6)	Wanted the account to be accessed easily
Easy to make	5.9% (29)	4.3% (10)	Password was easy to make
Habit	4.0% (20)	3.0% (7)	Did what they normally did for password creation
Avoid reuse	3.0% (15)	3.9% (9)	Specifically wanted to avoid reusing a password
Low-value account	3.0% (15)	2.2% (5)	Felt that the account is low-value or does not have sensitive information, and that influenced their choice of password
Recommendation	2.2% (11)	1.7% (4)	Chose their strategy because others recommended it
Avoid reset	1.8% (9)	2.6% (6)	Did not want to be troubled to reset the password
No need to remember	1.6% (8)	1.3% (3)	Strategy would obviate need to remember password
Frequent use	0.6% (3)	0.4% (1)	Account is used often
High-value account	0.6% (3)	0.4% (1)	Felt that this account is valuable/has sensitive information influenced password creation for this account
Speed	0.6% (3)	0.4% (1)	Wanted to access the service as quickly as possible
Easy reset	0.2% (1)	0.0% (0)	Resetting the password is easy
Fun	0.0% (0)	2.2% (5)	Following strategy is enjoyable to them personally
Just meet requirements	0.0% (0)	1.7% (4)	Minimally satisfies the account's password policy

Table 11: Motivations for participants' choice of password strategy (among 494 shared, 230 personal accounts). Participants sometimes indicated more than one motivation per password.

Digital Nudges for Access Reviews:
Guiding Deciders to Revoke Excessive Authorizations

Thomas Baumer
Nexis GmbH

Tobias Reittinger
University of Regensburg

Sascha Kern
Nexis GmbH

Günther Pernul
University of Regensburg

Abstract

Organizations tend to over-authorize their members, ensuring smooth operations. However, these excessive authorizations offer a substantial attack surface and are the reason regulative authorities demand periodic checks of their authorizations. Thus, organizations conduct time-consuming and costly access reviews to verify these authorizations by human decision-makers. Still, these deciders only marginally revoke authorizations due to the poor usability of access reviews. In this work, we apply digital nudges to guide human deciders during access reviews to tackle this issue and improve security. In detail, we formalize the access review problem, interview experts ($n = 10$) to identify several nudges helpful for access reviews, and conduct a user study ($n = 102$) for the *Choice Defaults Nudge*. We show significant behavior changes in revoking authorizations. We also achieve time savings and less stress. However, we also found that improving the overall quality requires more advanced means. Finally, we discuss design implications for access reviews with digital nudges.

1 Introduction

The Open Web Application Security Project (OWASP) lists "broken access control" as the Top 1 vulnerability and discovers it in 94% of the tested web applications [36]. Excessive authorizations are one driver for this OWASP vulnerability, as these are granted without an actual need and thus open an unnecessary attack surface. More precisely and within this paper, we ask highly qualified Identity and Access Management (IAM) experts to estimate the ratio of excessive autho-

rizations in mid- and large-sized organizations. Our experts expect about a fifth to a quarter of the authorizations to be excessive and vulnerable ($M = 22.8\%$, $SD = 6.4\%$, $n = 10$).

To mitigate this vulnerability, regulative authorities demand organizations to evaluate their authorizations with periodic access reviews. Well-known regulations include SOX [52], Basel III [6], MARisk [12], or HIPAA [51]. In large organizations, this involves hundreds of access review deciders for six figures of authorizations [18, 39]. These deciders (e.g., department heads) evaluate these authorizations within their responsibility. Although accountable, deciders face a time-consuming and frustrating task, as their expertise and objectives might not primarily match with security. Responsible deciders must also avoid mistakes: While revoked authorizations can interrupt their organization shortly, falsely confirmed excessive authorizations drive security risks [25]. Research [18] shows in a real-world case study that deciders only revoke 1.2% of the reviewed authorizations instead of the expected one-fifth excessive ones. Besides this clear need for improvement, only a few papers [18, 22, 26, 39] study access reviews.

As shown by Jaferian et al. [26], crucial issues for access reviews are rooted in poor usability. Using digital nudges to guide decisions [53] is thus a promising approach to improve access reviews. However, we identify several research gaps: First, current research does not formalize access reviews. Second, it is unknown how digital nudges address access review challenges. Third, it is unclear whether digital nudges actually improve access reviews. We investigate these research gaps with the following research questions:

Q1 *How to formalize the access review problem?*

Q2 *How do access review challenges map to digital nudges?*

Q3 *Does an applied digital nudge (the Choice Defaults Nudge) benefit the access review problem?*

This work follows a mixed methods approach in an exploratory sequential design. We use qualitative methods to define a formal and precise notation of the underlying problem (Q1) and to interview highly qualified experts ($n = 10$) about

applying digital nudges for access reviews (Q2). Moreover, our quantitative methods use insights of Q1 and Q2 to conduct a user study ($n = 102$) with an application of the *Choice Defaults Nudge* for access reviews (Q3). Consequently, our methods lead to the following contributions:

- We are first to formalize the access review problem.
- We map the expected effects of digital nudges to access review challenges based on 10 expert interviews. We find that access reviews benefit from digital nudges.
- We show behavior changes leading to quality improvements and more revoked authorizations by applying the *Choice Defaults Nudge* within a user study ($n = 102$). Moreover, we achieve time savings and lower frustration.

The remainder of this work is outlined as follows: Section 2 covers the background of this work, including relevant terminology, access review challenges, digital nudges, and related work. Section 3 provides details about our mixed method approach. Subsequently, we present the three-fold results of our paper. In Section 4, we first formalize the access review problem. Second, we map digital nudges with the access review challenges through the expert interviews in Section 5. Third, we conduct a user study on the *Choice Defaults Nudge* for access reviews in Section 6. Following the results, we discuss the general findings of this work in Section 7. Finally, Section 8 concludes and gives an overview for future work.

2 Background

2.1 Terminology

Identity and Access Management (IAM) is a cornerstone of modern cybersecurity, as it manages users and their access to sensitive data and services of organizations. Therefore, IAM provides tools to administer, authorize, and authenticate identities. Regulative authorities acknowledge the relevance of IAM and demand proper security controls. Besides state-of-the-art authentication, one crucial control is to demonstrate that the users still require granted access. Access reviews are a typical tool to prove the actuality of the granted access. These access reviews are the main focus of this work.

Access reviews are a periodic and compliance-driven process to verify users' authorizations. A team of domain experts, managers, application owners, and security admins typically reviews the granted authorizations with their knowledge of current processes, people, and resources. Especially in large organizations, access reviews are labor-intensive. Because of the recurring workload of access reviews, an organization might not finish an access review before the next one starts. The primary goal of access reviews is *revoking excessive authorizations*. Secondary goals are the determination of responsibilities for authorizations, requesting missing authorizations, or organization-specific data quality requirements. [18, 22, 26]

Nudges help humans make choices in analogous and digital systems. While these individuals must make their choices freely, *choice architects* design *choice architectures* to support their decisions by *nudging* towards a desired option. A nudge is thus a characteristic, influencing a decision in the interest of the decider. An example of a nudge in a supermarket is making healthy food easily accessible while making the unhealthy one harder to reach. From an ethical perspective, a nudge does not prevent a human from making a specific choice and only influences the decision in the best interest of the human. A *digital nudge* applies the idea of nudges to information systems. With features of user interfaces for guidance, users can make their choices freely and supported by the best advice of the choice architecture. [27, 45, 50]

2.2 Access Review Challenges

Based on expert interviews, Jaferian et al. [26] summarize access review challenges (C1-C5). We utilize these challenges throughout the paper, and thus detail them in this section.

C1: Scale outlines the number of involved IAM entities for the access review. The scale of the users, roles, permission, accounts, or assignments quickly grows into large numbers [18, 26, 39], making careful considerations for organizing the access review's workload necessary. Furthermore, the heterogeneity of these entities within real-world organizations intensify this challenge [30].

C2: Lack of Knowledge refers to the understandability of roles and permissions [26, 30, 31]. IAM entities might not have telling names, comprehensive descriptions, or concepts like roles or permissions might not have been fully understood. Experts thus might take uninformed or *best guess* decisions, leading to a bias for keeping unnecessary granted authorizations, violating the Principle of Least Privilege (PoLP) [18]. Additionally, for large organizations, the knowledge about these entities is distributed (or even missing completely), making the advice of responsible domain experts necessary [30].

C3: Frequency describes a dilemma for the managers: access reviews are not their *actual* responsibility, but they are frequently asked for it [26]. The experts might not feel a need to participate, leading to failing access reviews. Ultimately, this may cause even more access reviews, since successfully executed access reviews are part of compliance and audits. Thus, while access reviews usually are only required yearly, some organizations execute them quarterly, hoping not to fail access reviews due to lack of participation [26].

C4: Human Errors are common due to the scale and manual execution of access reviews by human deciders. These experts ultimately decide about required or excessive access by applying the best of their knowledge. This process is, therefore, inherently error-prone, as decisions to the best of the experts' knowledge might be incorrect or uninformed [18, 26].

C5: Exceptional Cases occur due to the scale and complexity of access reviews. Context knowledge is sometimes

required for an informed authorization decision. For example, some members of organizations might replace others while on leave, trainings or tests might require temporary access, etc. might cause disturbances [26].

2.3 Digital Nudges

Based on a literature survey, Jesse and Jannach [27] propose a taxonomy for digital nudges. The authors distinguish four primary categories with further sub-categories of digital nudges (N01-N13): decision information (N01-N04), decision structure (N05-N08), decision assistance (N09-N10), and social decision appeal (N11-N13). We refer to these nudges throughout the paper, and thus explain them in this section.

Decision Information tries to present information helpful for the decision-maker without altering the available choices. This category comprises information translation (N01), salience (N02), visibility (N03), and phrasing (N04).

- *N01: Information Translation* targets reducing the cognitive effort for a decision by simplifying information or decreasing vagueness and ambiguity [48].

- *N02: Information Salience* aims to raise or decrease the prominence of information, by visualizations or making information harder or easier to notice [11, 48].

- *N03: Information Visibility* fosters decision information. This category includes mechanisms to disclose [24, 28, 48], compare [11, 48] or warn with [24, 33, 48] (tailored [28, 33] or external [35, 49]) information.

- *N04: Information Phrasing* puts presented information in context to intervene with the decisions to make. This category comprises the utilization of heuristics or biases like anchoring [33, 34, 48], availability [44, 50], the endowment effect [11, 44], framing [11, 33, 48], loss aversion [34, 44, 48], priming [11, 48, 50], etc.

Decision Structure alters the decision arrangement, comprising the decisions' range & composition (N05), defaults (N06), consequences (N07), and required effort (N08).

- *N05: Range & Composition* groups and categorizes choices. Therefore, choice architects or the decision-makers themselves break large decision structures into smaller category partitions [28, 48, 53], to present these one after another [28, 33] or to make them more comparable to each other [28, 35]. Choice architects can also utilize ordering effects for the presented options [11, 48].

- *N06: Choice Defaults* is one of the most effective and well-studied nudges [24]. The nudge preselects choices without hindering the decision maker from actively making another choice. On the one hand, a decision-maker is more invested in an actively made decision [35, 48]. On the other hand, decision-makers

rather accept the preselected status quo than actively decide against it [11, 24, 35, 48].

- *N07: Option Consequences* add further yet rational insignificant effects to the choice without changing the overall economic incentives. These consequences include social outcomes or minor benefits & costs [24, 35].

- *N08: Option-related Effort* modifies the effort or ease to make decisions. This nudge includes capping [11, 48] or raising financial & physical effort [24, 35] of decision-makers choices to mitigate mindless actions. Furthermore, eased and more convenient choices speed up decisions, e.g., making desired choices more accessible [48].

Decision Assistance aids decision-makers to realize their intentions. This category includes the usage of reminders (N09) and commitment facilitation (N10).

- *N09: Reminders* actively put already available information into or out of the attention focus of the decision-makers. This nudge includes reminding of underlying goals, deadlines, and their relevance [11, 24, 33, 35, 48, 49] or stating social expectations for decisions [35].

- *N10: Commitment Facilitation* helps decision-makers to (timely) finish their asked for decisions. This nudge includes precommitment strategies (e.g., user-defined sub-goals) [24, 33, 35, 48] or public commitment (e.g., pressure by publicly communicating own goals) [11, 35].

Social Decision Appeal category focuses on the social implications of nudges, including the Messenger Reputation (N11), Social Reference Point (N12), and Empathy Instigation (N13).

- *N11: Messenger Reputation* considers the reputation of the messenger delivering the information for the nudge. On the one hand, the messenger effect nudges a decision-maker since the messenger provides a certain and influencing impression about itself. For example, an actually well-designed and important choice architecture might dilute its seriousness if it contains many spelling mistakes [44]. On the other hand, the reputation of a system can be improved when choice architects expect and forgive the errors of their decision-makers [28, 53].

- *N12: Social Reference Point* nudges a decision based on social opinions. E.g., the opinion of a majority (Argentum-Ad Populum), group (Group-Ad Populum) [16], or an opinion leader [35] can influence decision-makers. Additionally, deciders tend to follow a herd [34, 44, 48] and might desire a comparison with their peers influencing their own decisions [24, 33, 35].

- *N13: Empathy Instigation* uses feelings to influence deciders. For example, an avatar might smile or cry upon the choices of a decision-maker (moral suasion) [11, 48], or a choice architect can trigger reciprocity by doing something good for the decision-makers to nudge them into returning the favor with good choices [11].

2.4 Related Work

Access control ensures users can only act within their intended authorizations and is characterized by its necessary yet cumbersome maintenance. Related research on maintenance covers more efficient access control models, optimization, and general maintenance processes like access reviews. By evolving from access control matrices [41], the most dominant access control models are Role-Based Access Control (RBAC) [15, 37, 43] and Attribute-Based Access Control (ABAC) [23, 46] as these reduce maintenance costs. Modeling access control policies considers bottom-up, top-down, or hybrid approaches [14] but often overlook their actual optimization without recalculating them [31, 38]. Therefore, access control maintenance targets up-keeping authorizations in changing needs and environments based on IAM goals [25, 30]. This includes periodically reviewing and revoking excessive access [18, 22, 26], granting missing access [47, 54], and timely propagation [7] to maintain secure authorizations. This paper especially relates to work on maintenance by access reviews: Jaferian et al. [26] study its challenges and usability. Puchta et al. [39] show positive effects on using external data for access reviews. Groll et al. [18] assess decision quality. Hill [22] conducts a case study for HIPAA [51] compliant access reviews.

Digital nudges are a popular research topic, as shown by various surveys: While Bergram et al. [9] conduct a general literature review, Schaer and Stanoevska-Slabeva [44] analyze digital nudges in customer-journeys and Jesse and Jannach [27] in recommender systems. Additionally, a survey of Caraban et al. [11] covers a practical and ethical application. As an established means to shape human behavior, applications of (digital) nudges exist for many domains. Examples include e-commerce [2, 13], sustainable smart home [8], contract tracing [17], or cybersecurity. In detail, cybersecurity examples include digital nudges to prevent phishing [55] or increase password quality [29, 56, 57]. An application of digital nudges for access reviews remains open so far.

3 Methods

This research uses mixed methods in an exploratory sequential design. First, we use qualitative methods to formalize the Access Review Problem (ARP) (Q1) and relate access review challenges to digital nudges (Q2). Second, we use these qualitative insights in quantitative methods to study the effect of the *N06: Choice Defaults* for access reviews (Q3). Third, a discussion wraps up the findings. Figure 1 depicts our mixed methods. In the following, we detail each part.

3.1 Formalizing the Access Review Problem

While the access review challenges comprise a global view, we formalize the actual Access Review Problem (ARP) in

Section 4. Its goal is to understand the underlying problem better. This formalization targets a quantifiable and comparable foundation for the solution of the ARP. Thus, we argue access review as a transition between two authorization states, depicted as confusion matrices. This precise formalization of the ARP is the basis for the hypotheses of the user study.

3.2 Relation of Access Review Challenges to Beneficial Digital Nudges

Complementary challenges to the ARP are discussed in the literature, including scale, lack of knowledge, frequency, human errors, and exceptional cases [26]. Digital nudges are a promising approach to address the ARP and its challenges. But it is unknown, whether digital nudges can help and which effects can be expected from their application (Q2).

To better understand this relationship between access review challenges and digital nudges, we investigate and map access review challenges from Jaferian et al. [26] with the digital nudge taxonomy of Jesse and Jannach [27] by conducting semi-structured expert interviews based on the guidelines of Adams [1]. The interviewed industry experts provide practical experience in access control and reviews. Therefore, we target highly qualified professionals with at least five years of experience working with large IAM systems, periodical executed access reviews, and managing thousands of identities or consultants with practical experience for many enterprises. Of course, these highly qualified experts are not readily available, but we managed to acquire 10 of these experts through personal and professional contacts. The experts are located in Germany. We use their expertise for a well-grounded argumentation for the relationship between access review challenges and digital nudges. Section 5.1 details further on the method for the expert interviews.

3.3 User Study for the Choice Defaults Nudge

After laying out theoretical foundations for digital nudges and access reviews in Sections 4 and 5, we study the application of a selected digital nudge in-depth. The expert mapping of digital nudges and access review challenges suggests several digital nudges. To sharpen the scope of the use study, we select *N06: Choice Defaults* based on the following reasons:

- Literature considers *N06: Choice Defaults* among the most effective digital nudges [24].

- The expert interviews had strong positive and negative expectations, inviting a more detailed examination.

- We felt confident to apply the *N06: Choice Defaults* to an access review and study its effects precisely.

We design an access review, simulating a real case: experts often describe access reviews as repetitive, time-consuming, and tedious tasks, requiring a strenuous thought process to
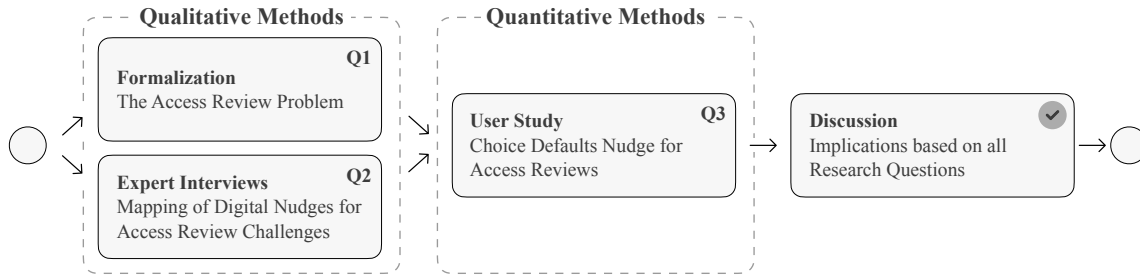
Figure 1: Mixed methods approach for this study.

determine correct authorizations. We thus hand out each participant a one-pager about the case. Participants manage a fictitious marketing department containing three teams within the case: graphic design, social media, and event management. The instruction describes the functions and tasks of each team and explains the unwanted implications of excessive or missing authorizations. While it is theoretically possible to review each decision using the document correctly, it takes some thought to make a correct decision.

To study the *N06: Choice Defaults* in-depth, we use three distinct configurations for access reviews with the same data basis: default accept, default reject, and a neutral default. This directly compares the default accept and reject configuration with a neutral state. The default accept configuration preselects every decision with an accept, the default reject vice versa, and the neutral default does not preselect.

We acquire 102 participants from a university context in Germany and randomly assign them to one of the three *N06: Choice Defaults* configurations. We select our sample size based on similar papers (c.f. Caine [10], also for the expert interviews). The (under-)graduate students have mostly a background in business informatics and IT security, indicating that they know essential IT security concepts and enterprise information systems. Furthermore, the participants are unaware of the research objective on digital nudges. We raffle a €100 gift card to one lucky participant to motivate participation. The participants must log in with authenticated accounts to avoid repeated participation and enable remote participation. We pilot the study with fellow researchers. Section 6.1 provides further details for the method.

3.4 Ethical considerations

Our experts were informed and consented to an anonymous publication of parts of their interviews. We will not share the recordings and delete the data one year after the publication.

The Institutional Review Board (IRB) *German Association for Experimental Economic Research e.V* approved the user study to comply with ethical requirements for working with humans. The certificate is available online.[1]

---

[1] https://gfew.de/en-ethik/HQwmKGTZ

4 Q1: The Access Review Problem

To better understand access review and benchmark our user study design, we introduce a representation of granted authorizations and security policies within a confusion matrix depicting User Permission Assignments (UPAs). Figure 2 maps the actually granted authorizations with security policies. We assess authorizations as effective access grants (which may contain errors), while security policies define the conceptual access users should have (ground truth). We construct a classical confusion matrix by mapping these authorizations and security policies with a binary distinction. Thus, the effectively granted UPAs are Predicted Positive (PP) as $PP = TP + FP$, while $P = TP + FN$ should be granted. *PN* and *N* are vice versa not-granted UPAs. Therefore, the True Positives (TPs) describe UPAs, granted in reality and conceptually. The sensitivity $SEN = \frac{TP}{P}$ represents the rate of correctly granted UPAs. Vice versa, True Negatives (TNs) describe UPAs, not granted in reality and in concept. The specificity $SPC = \frac{TN}{N}$ represents the rate of correctly not-granted UPAs. Together, sensitivity and specificity express the balanced accuracy $BA = \frac{SEN + SPC}{2}$, equaling 100% in a perfect world without errors.

		Authorization	
		Positive *PP*	Negative *PN*
Security	Positive *P*	*TP*	*FN*
Policy	Negative *N*	*FP*	*TN*

Figure 2: Confusion matrix for UPAs.

However, type I (False Positives (FPs)) and type II (False Negatives (FNs)) errors are present in reality. On the one hand, FPs are granted authorizations not considered by security policies (excessive UPAs). These excessive authorizations drive security risks, as over-privileged users are a target for threat actors. The primary goal for access reviews is lowering FP, which is highlighted in Figure 2. On the other hand, FNs are mistakenly not granted authorizations (missing UPAs). An example of their impact is when users cannot do their legitimate tasks because they do not have access to the required systems. This causes dissatisfaction for the users and slows down processes. In a relative notation, the False Discovery Rate (FDR) describes the percentage of excessive UPAs FP

based on PP as $FDR = \frac{FP}{PP}$. Vice versa, the False Omission Rate (FOR) describes missing UPAs as $FOR = \frac{FN}{PN}$.

Thus, an access review can be understood as transitioning from one UPA set depicted as confusion matrix C^1 to another C^2. The primary goal is to reduce the FDR while retaining or improving BA. We introduce definitions for Access Reviews (ARs) and the Access Review Problem (ARP) as:

Definition 4.1 (Access Review (AR)). Given a confusion matrix C^1 describing an UPA^1 set, an access review AR revokes a subset of the effectively granted authorizations $R \subset PP^1$. When executing AR a confusion matrix C^2 describes the resulting set as $UPA^2 = UPA^1 \setminus R$.

Definition 4.2 (Access Review Problem (ARP)). Design AR in such a way that a (human) deciders can review and revoke UPAs $R \subset PP^1$ according to their knowledge about security policies P^1, that the FDR is reduced ($FDR^1 > FDR^2$), without lowering BA ($BA^1 \leq BA^2$). The ARP is solved on a $FDR^2 = 0\%$ without decreasing BA: $BA^1 \leq BA^2$.

The following hypotheses hence allow testing whether an access review design improves the ARP:

H₀ An access review design does not improve the ARP as the FDR remains or rises $FDR^1 \leq FDR^2$ or BA remains or decreases $BA^1 \geq BA^2$.

H₁ An access review design improves the ARP as the FDR decreases $FDR^1 > FDR^2$ and the BA raises $BA^1 < BA^2$.

5 Q2: IAM Experts on Digital Nudges

5.1 Method Details

The interviews comprise three phases: an interviewee introduction, an explanation of access review challenges and digital nudges, and a workshop to generate the mapping of access review challenges and digital nudges. (i) The interviewee's introduction collects data about their access review experience, their perspective on its challenges, and their estimation of excessive authorizations (FP). (ii) The explanation phase ensures essential knowledge about digital nudges, reminds the interviewee of access review challenges, and ensures a common vocabulary. We use the interviewees' perspectives on access review challenges to explain to them the access review challenges of Jaferian et al. [26]. (iii) The procedure for querying the mapping for each considered digital nudge [27] follows this scheme: First, we explain the digital nudge in general and provide a suitable example for the interviewee. Afterward, we let the expert freely reflect on the effect of this digital nudge and its benefit to all access review challenges. Finally, we ask the expert to rate each access review challenge on a five-level Likert scale from very positive (+2) to very negative (-2). This rating scheme helps the expert to express

their arguments more comparable to each other. The complete interview script is available in Appendix A.1.

We interviewed 10 highly qualified experts with experience in conducting several Access Reviews (ARs) specialized for IAM by implementing IAM tools (engineers), responsible for managing thousands of users in IAM systems (inhouse), or advising clients (consultants). Table 1 protects their identities but depicts their high expertise for ARs. The interviews took an average of 60 minutes and were recorded, transcribed, coded, and evaluated. We translated relevant parts of the interviews into English during the coding process.

Table 1: Participants for expert interviews.

Interview	Experience				Sector
	Years	Clients	Users	ARs	
E01: IAM consultant	8	40		20	Multiple
E02: IAM consultant	5	15		10	Multiple
E03: IAM engineer	12	40		15	Multiple
E04: IAM inhouse	8	15	1k	40	Insurance
E05: IAM consultant	19	25		10	Multiple
E06: IAM consultant	13	40		25	Multiple
E07: IAM consultant	6	15		50	Multiple
E08: IAM inhouse	15	2	19k	4	Biotech
E09: IAM consultant	11	4		10	Banking
E10: IAM inhouse	7	1	13k	120	Insurance

We recorded the interviews with Microsoft Teams, transcribed them with Word, and summarized and coded them with Excel. For the coding, we use both deductive and inductive coding [3]. Since we already know the access review challenges [26], we first applied deductive coding based on these challenges for each digital nudge. This deductive coding already sorts large parts of the interviews in proven codes. However, we noticed that several augmentations exist within these codes. Thus, we also developed inductive codes for each nudge and challenge combination to capture the interviews comprehensively. For the rating of each nudge and challenge pair, we initially used the mean expert ratings. After coding and comparing the interviews, we slightly adapted the ratings, to balance well-reasoned arguments across the experts. The resulting codebook is available in Appendix A.2.

5.2 Results

This section presents the experts' mapping. We build on the presented background of the access review challenges (C1-C5) in Section 2.2 and digital nudges (N01-N13) in Section 2.3. The resulting mapping is depicted in Table 2, whereas the challenges serve as columns and the digital nudges as rows. The cells summarized a rating for each challenge and nudge. In the following, we detail each digital nudge.

N01: The experts stress the benefits of comprehensible data. While C1 and C3 do not decrease, comprehensible data indirectly increases its learnability and comfort for the deciders, easing management eventually. For C2 and C4, the

Table 2: Nudges [27] and access review challenges [26].

Nudges	C1	C2	C3	C4	C5
N01: Information Translation	1	2	1	2	0
N02: Information Salience	1	0	1	1	2
N03: Information Visibility	1	2	0	1	2
N04: Information Phrasing	0	-1	0	1	0
N05: Range & Composition	2	1	1	2	2
N06: Choice Defaults	2	-2	2	-2	0
N07: Option Consequences	0	-1	1	-1	-1
N08: Option-related Effort ↗	-1	1	-1	1	1
N08: Option-related Effort ↘	1	-1	1	-1	-1
N09: Reminders	0	1	2	-1	0
N10: Commitment Facilitation	1	0	1	1	0
N11: Messenger Reputation	1	2	1	2	2
N12: Social Reference Point	0	2	0	1	2
N13: Empathy Instigation	1	1	1	1	0

Note: Option-related effort is ↗ = increased, ↘ = decreased. The Likert scale spans from very positive +2 to very negative -2.

experts anticipate a strong positive effect, as comprehension is essential for C2: *"If data is displayed more comprehensibly, it's helpful for users with little knowledge [C2] about the decision." (E06)*. Being comfortable with the data is relevant for C4: *"If the user is comfortable with the displayed data, you can expect fewer human errors [C4]." (E07)*

N02: The experts emphasize the focus: *"In my opinion is the highlighting of C5 the only option to manage large data sets." (E05)* However, *"it depends on the quality of the highlighting" (E03)*, since excessive or missing highlighting might draw away attention from relevant decisions. But upon sufficient and reliable quality, decision-makers can efficiently focus on the highlighted decisions or attributes and decide the remainder quicker (benefit for C1 and C3). Decision-makers *"actually want to decide diligently but are hindered by its scale. These decision-makers could diligently and mindfully decide just the highlighted decisions in an efficiency tradeoff." (E09)*

N03: Showing additional data is crucial for C2 and C5 to make informed decisions while streamlining the focus to relevant attributes (C4). By only offering limited attributes for each decision in default, the management of C1 is eased. However, the user might not know the relevancy of specific hidden attributes as these move out of focus (C4).

N04: Our interview partners express reservations as decisions might not be based on rational knowledge but on biased phrasing (C2). However, for a well-executed implementation, its utilization can raise the access review acceptance (C4) as its relevancy could be communicated more effectively.

N05: The setup of meaningful partitions and sorting imposes overhead compared to just showing all decisions in one turn, thus worsening C1 and C3. However, the experts anticipate quite positive effects on all challenges. Similar sorted or clustered partitions leverage efficiencies as decisions transfer to whole partitions. These efficiencies ease the management for C1 and C3 since the workload decreases, while more consistent and mindful decisions mitigate C2 and C4. Furthermore, clustering and appropriate communication of exceptional cases (C5) can positively influence.

N06: The experts discuss the strong effects of *N06: Choice Defaults*. Due to the reduced workload by the preselection, the experts rate a positive effect on C1 and C3. However, the experts worry that deciders adopt a preselected default without further thought, leading to uninformed (C2) and mindless (C4 and C5) decisions. While a mindful default prevents errors on uncertainty (like for C5) or on evident cases, just adopting the recommended default can become a fallacy, assuming the recommendation algorithm's imperfections. This is especially an issue if the decision-makers trust the preselection so that they mindlessly adopt the default instead of a mindful decision. A falsely set default would then lead to a systematic bias, endangering the next audit relevant to compliance. In sum, the experts anticipate the potential of *N06: Choice Defaults*, but advise careful application.

N07: *"In practice, negative consequences dominate. For example, we will tell your boss if you don't finish your access review tasks within 14 days." (E01)* The experts acknowledge that creative and positive consequences could be feasible and reasonable, making frequent access reviews more comfortable (C3). However, they doubt there would be a game-changer in the long term because the effects would wear down over time (C3), and the decisions might be based on avoiding pressure or pursuing benefits (C4) instead of reason (C2 and C5). In this context, it is also worth noting that *"disadvantaged individuals need special consideration" (E09)* because finishing an access review in time might not be fair for these (C5).

N08: This nudge's influence on the access review challenges is ambivalent, as it depends on whether the option-related effort is increased (↗) or decreased (↘). If the effort *increases* (vice versa for decrease), the users take more time to decide. For C1 and C3, this worsens the situation as the workload rises with its time consumption. Taking more time for a decision (e.g., requiring a reason for confirming a high-risk authorization) also benefits C2, C4, and C5, as the decider would need to consider a reason or reconsider the decision. But the experts also stressed the efficiency and acceptance of the access review, as some users easily become annoyed by increased effort: E.g., *"We once required the users to set a note for the reviewed authorizations, but one user just put question marks for every note to bypass the input check." (E04)*

N09: *"By a simple reminder [email], we observe more participation." (E10)* While reminders are especially relevant for C3 to communicate open tasks or instructions and goals for access reviews (C2), they can also pressure decision-makers to decide quickly but uninformed (C4). The audience and channel of reminders are also essential for C4. E.g., an inexperienced decider might require instructions or training. The experts also noted that reminders via an email channel

dominate in practice but are quickly perceived as spam. *"Everybody wants something from all colleagues. Ironically, some colleagues even configure automated email filters which they won't check afterward." (E09)* In this sense, a personal or multichannel address is most effective, but it is a considerable effort for the IAM team conducting the access review.

N10: The experts appreciate the autonomic commitment in combination with semantic partitioning (N05) of the decisions. An autonomic configuration of sub-goals and sub-deadlines suitable for the deciders benefits C1 and C3 as the deciders *"perceive control over scale and frequency" (E06)*. This leads to more comfort, as sub-goals and sub-deadlines become meaningful for the deciders, mitigating C4.

N11: The experts stress the importance of this nudge: *"Most important point; If the IAM team is not accepted, it is going be tough." (E04)* Furthermore, they note its failure in practice: *"Access reviews are usually perceived negatively." (E10)* With a suitable messenger reputation, users will trust and endure the tedious tasks of the access reviews more, which is beneficial for C1 and C3. The experts also anticipate strong benefits for C2, C4, and C5 as the decision-makers will dare to ask or tell an approachable IAM team their relevant questions or mistakes: *"If the IAM team is approachable, users communicate errors more eagerly or at all." (E07)*

N12: Similar to N11, if the social reference point sympathizes with the access review, decision-makers are likely to endure the tedious workload (C1 and C3). However, on low sympathy, the opposite effect might apply. The experts anticipate positive effects for C2, C4, and C5 because deciders discuss the access review: *"For example, we introduced access review chat groups for business units. Decision-makers can talk about access reviews, like showing their own or seeing others' progress, asking questions, etc." (E07)* In this sense, exceptional cases (C5) might become evident after a discussion and sharing knowledge about similar cases (C2), while noticing the colleagues' progress might remind stragglers or expose them to peer pressure (C4).

N13: *"On large scale [C1] and high frequency [C3], the decision-makers want to work with a pleasant tool." (E06)* Moral suasion and empathetic feedback (C2) can inform and convince the decision-maker about odd user behavior (e.g., mindlessly accepting all authorizations) without losing their motivation (C4). Reciprocity also fosters mitigation of C4 by *"always addressing the positive side: the access review is meant to help you, the decision-maker, to compliantly and securely maintain your authorizations." (E08)*

Furthermore, the experts estimate a mean on excessive authorizations (FP) at 22.8% ($SD = 6.4\%$). Since we also asked our experts about common AR challenges, we confirm the AR challenges first published by Jaferian et al. [26].

In summary, our experts conclude positive and negative effects when using digital nudges. Table 2 summarizes these key takeaways. We hope to motivate future work with it as most digital nudges invite dedicated research on access reviews.

6 Q3: Choice Defaults in Access Reviews

6.1 Method Details

In the data set of the user study (Appendix B.1), we let participants review (accept or remove) granted UPAs $PP = TP + FP$ (legit TP or excessive FP), leading to UPA revoke operations only. Not granted UPAs $PN = FN + TN$ (missing FN or legit TN) are not considered. After piloting, we determined 160 UPAs serving as decisions to align an estimated study duration of 20-30 minutes and not to deter participation. Therefore, the crafted data set comprises 160 UPAs (PP) split into 80 legitimate ones (TP) and 80 excessive ones (FP), clearly distinguished by a case study document (see Appendix B.2). Figure 3 summarizes the initial UPAs as a confusion matrix.

		Authorization	
		$PP = 160$	$PN = 232$
Security	$P = 80$	$TP = 80$	$FN = 0$
Policy	$N = 312$	$FP = 80$	$TN = 232$

Figure 3: Confusion matrix for the case of the user study.

We configure and execute the access reviews with the commercial tool NEXIS4[2]. The tool can import our data set, configure *N06: Choice Defaults*, execute large-scale access reviews, and collect relevant data points. Figure 4 displays a simplified screenshot of the review process. Further screenshots for all groups are available in Appendix B.3. For data collection, we make three observations for each access review participant: their decisions for the 160 UPAs, their time consumption, and their self-assessment for the NASA Task Load Index (TLX) [20]. (i) The tool stores each binary decision out-of-the-box, leading to a total of 16,320 manual decisions for 102 participants and 160 UPAs. (ii) We measure the time consumption for each participant by comparing the events for starting the access review and confirming the final completion prompt. (iii) After completion, we ask the participants to fill out a questionnaire for the NASA TLX [20] to capture their perceived workload. These questions are based on a Likert scale (-3 to +3) and include:[3]

- Mental Demand: How mentally demanding was the task?

- Temporal Demand: How hurried or rushed was the pace of the task?

- Performance: How successful were you in accomplishing what you were asked to do?

- Frustration Level: How insecure, discouraged, irritated, stressed, or annoyed were you?

---

[2] https://nexis-secure.com/en/
[3] We omitted the questions for physical demand and effort, as these are not applicable or relevant for our study.

Figure 4: Simplified screenshot of the access review.

During the post-processing of the study, we used Microsoft Excel and R[4] for data cleansing or data analysis. Data cleansing primarily comprises capping the time consumption for the AR to 60 minutes, as some participants took a break. We calculate the means, standard deviations, and non-parametric ANOVA of the AR confusion matrix, time-consumption, and NASA TLX indices. For our exploratory analysis of correlations (Spearman) and local regressions, we utilize a pair plot generated in R (see Appendix Figure 9). Supporting the open data idea, we publish all data to replicate our results on GitHub: https://github.com/AccessReview/Availability.

6.2 Results

This Section summarizes our observations of the user study (see Table 3). A post-hoc power analysis based on ANOVA for our three groups ($n = 34$) and an $\alpha = .05$ results in effect powers of .13 for a small effect ($f = .1$), .6 for a medium effect ($f = .25$), and .95 for a large effect ($f = .4$).

For all 102 participants, the mean review time t for the 160 decisions is $t = 22$ minutes with $SD = 13$ minutes. Deciders of all groups used to over-accept authorizations, amounting to a total accept rate of $1 - \frac{R}{PP} = 56.1\%$ (rather than a $SEN = 50\%$). H_0 is rejected for 99 of 102 reviews. The remaining 3 participants failed to achieve an ARP improvement. All participants' mean BA increased from 87.2% to 91.2% ($SD = 7.9\%$). The false discovery rate FDR, which represents the amount of excessive authorizations, was reduced from 50.0% to 21.6% ($SD = 14.7\%$). This improvement came at the cost of some erroneous revokes, leading to a mean FOR of 2.9% ($SD = 3.5\%$). In sum, most participants improved the ARP. The result data shows that two deciders behaved as "spammers" by either blindly accepting all authorizations (one decider in the accept group) or blindly rejecting them all (one decider in the reject group). These participants are among the three who failed to improve the ARP. While the data set is too small to make this finding statistically significant, it seems evident that the spammers just adopted the default.

The neutral configuration group is a control group for the default accept and reject nudge. Users from this group took a mean time of $t = 26$ minutes ($SD = 15$) and accepted 57.8% of the authorizations. The neutral group estimated the temporal demand as slightly low, with a mean score of -.8. On average,

neutral users stated the mental demand to be slightly high (.9) and their frustration to be neutral to slightly high (.5). They estimated their performance to be slightly above average (.9). The achieved BA is 91.9% ($SD = 5.8\%$), with the error rates FDR of 21.0% ($SD = 10.7\%$) and FOR of 2.6% ($SD = 2.5\%$).

The accept group only took $t = 19$ minutes ($SD=10$). With a time save of 24.3% to the neutral group. While the perceived TD was unchanged at -.8, both FL and MD were reduced by almost one point to a score of -.2 ($\Delta = -.7$) and .2 ($\Delta = -.7$). The accept rate was slightly higher than in the neutral group with 58.7% (+.9%). The default accept group achieved a BA of 92.3% ($SD = 5.3\%$), scoring .4% higher than the neutral one. The error rates were also marginally better than in the neutral group with $FDR = 20.8\%$ ($\Delta = -.2\%$, $SD = 9.3\%$) and $FOR = 2.2\%$ ($\Delta = -.4\%$, $SD = 2.6\%$).

Like the accept group, deciders of the reject group finished quicker than the neutral group with $t = 21$ minutes ($\Delta = -16\%$, $SD = 13$). Again, the estimated TD of -.4 did not reflect this ($\Delta = +.4$), but the stated FL and MD were reduced to -.6 ($\Delta = -1.1$) and -.2 ($\Delta = -1.1$). Unlike the accept group, however, the reject group showed a considerably reduced accept rate of 51.8% (-6.0%), which is very close to the initial $SEN = 50\%$. Unfortunately, the increased willingness to revoke did not improve the results: The deciders revoked fewer excessive authorizations than the neutral group ($FDR = 22.9\%$, $\Delta = +1.9\%$, $SD = 21.4\%$) and more correct ones ($FOR = 3.9\%$, $\Delta = +1.3\%$, $SD = 4.7\%$). With $BA = 89.4\%$ ($SD = 11.2\%$), BA was still improved regarding the initial state ($\Delta = +2.2\%$), but worse than the neutral configuration ($\Delta = -2.5\%$).

We ran a non-parametric Kruskal-Wallis test ($\alpha = .05$) to check for the significance of our observations between the three groups. We detect differences for the number of revokes R ($p = 0.039$), indicating that *N06: Choice Defaults* did affect users' willingness to accept or reject authorizations. We also confirm differences for MD ($p = .049$) and FL ($p = .038$), indicating that lower stress perceptions result from the applied *N06: Choice Defaults*. We used Dunn's test for pairwise comparisons, showing that the neutral and reject groups differ for MD ($p.adj = .045$) and FL ($p.adj = .038$). However, the quality metrics BA, FDR, and FOR did not differ significantly between the study groups, which is unsurprising since the data set balances TP and FP at 80.

A test for Spearman correlation showed no significant correlation between review duration t and any of the quality metrics (BA, FDR, FOR), indicating that quality did not depend on the time spent. The data shows a significant positive correlation between the deciders' frustration level FL and t for the total population (.286) and the neutral group (.403), as well as between the stated mental demand MD and t (total: .237, neutral: .423). FL and MD are strongly correlated for all groups (total: .646, neutral: .589, accept: .672, reject: .664). We follow that deciders did not strictly distinguish between MD and FL and that longer reviews are perceived as more frustrating and/or mentally demanding. Interestingly, the per-

---

[4]https://www.r-project.org/

Table 3: General summary of the user study, including arithmetic means and standard deviations.

Group	n	Fails	t		R		BA		FDR		FOR		MD		TD		PF		FL	
			M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Initial	-	-	-	-	-	-	.872	-	.500	-	.000	-	-	-	-	-	-	-	-	-
Total	102	3	22	13	70.3	19.2	.912	.079	.216	.147	.029	.035	.3	1.6	-.6	1.5	1.1	1.6	-.1	1.7
Neutral	34	0	26	15	67.5	12.4	.919	.058	.210	.107	.026	.025	.9	1.4	-.8	1.5	.9	1.5	.5	1.7
Accept	34	1	19	10	66.1	19.5	.923	.053	.208	.093	.022	.026	.2	1.5	-.8	1.5	1.2	1.6	-.2	1.8
Reject	34	2	21	13	77.2	22.8	.894	.112	.229	.214	.039	.047	-.2	1.8	-.4	1.6	1.1	1.7	-.6	1.6

Note: M for arithmetic mean and *SD* for standard deviation. *n* for the participant count and *Fails* for executions in rejecting H_0. *BA*, *FDR* and *FOR* for measuring the Access Review Problem (ARP). *t* for the time-consumption of the AR and *R* for the amount of rejected UPAs. *MD* (Mental Demand), *TD* (Temporal Demand), *PF* (Performance) and *FL* (Frustration Level) for the NASA TLX.

ceived temporal demand *TD* did not correlate with *t*, possibly due to a missing baseline of a "normal" review duration. The result data showed a strong positive correlation between the perceived performance *PF* and actual performance *BA* (total: .607, neutral: .635, accept: .605, reject: .639), and a negative one between *PF* and the error rates *FDR* (total: -.336, neutral: -.516, reject: -.422; accept: not significant) and *FOR* (total: -.541, neutral: -.568, accept: -.480, reject: -.603). Therefore, the deciders had a realistic estimation of their performance. The result data also showed significant negative correlations between *FL* / *MD* and *BA* as well as positive ones between *FL* / *MD* and the error rates *FDR* / *FOR*, each for some groups. However, the causality remains unclear if deciders who find the task more difficult experience more stress, more stressed deciders deliver poor results, or both. Figure 9 (Appendix) shows the Spearman correlation and the local regressions.

➤ **Key takeaways of our user study:** (i) Almost all deciders improved the ARP. (ii) The required time differed substantially but was unrelated to quality (BA). (iii) *N06: Choice Defaults* led to reduced time effort and stress perception. (iv) A default reject led to more rejects. (v) A simple *N06: Choice Defaults* did not affect quality (BA) significantly but influenced the number of rejects. In detail, however, some increase in false rejects is tolerable as false accepts legitimate excessive authorizations leading to a false sense of security. (vi) Deciders' self-assessed performance correlates significantly with BA, indicating the deciders' realistic self-assessment.

7 Discussion

7.1 Acceptance Bias

Participants of the user study tend to accept existing authorizations. Existing research already documents and analyzes over-granting in real-world scenarios [18, 47, 54]. However, such scenarios involve strongly imbalanced data (see expert interviews: $1 - 22.8\% = 77.2\%$ of authorizations are estimated to be correct), social implications (a revoke acts against the interests of a real person), and unequal visibility of the two error types. Erroneous revokes are detected quickly, and the decider alone is responsible, while erroneous accepts are not

immediately visible and all previous approvers share the responsibility for also not resolving the error. With an initial *SEN* of 50% and no personal repercussions, the study had none of these biases and made no implication that acceptance is favorable to revocation. Still, deciders accept authorizations too often, with an average accept rate of 57.8% in the neutral group (see Section 6.1). While the study data does not explain this behavior, a possible explanation might be that the status-quo bias discourages deciders from revoking [42]: Following a real-world scenario, the study description states that participants need to review *existing* authorizations, which would be revoked upon rejection. The existence of a general status quo bias could also explain the relatively weak effect of the default accept bias on the accept rate: Study participants with default accept or reject nudge configuration needed to change an existing preselection to make an active decision and are thus also confronted with a status quo bias. If a status quo bias is already the reason for over-accepting in the neutral group, the effect of the default accept nudge would only repeat an already present bias. In contrast, the default reject nudge creates a new status quo that nudges the deciders in the opposite direction. The explanation seems plausible based on the study results, as the accept rate of the default accept group is closer to the neutral group (58.7%), and the accept rate of the default reject group is closer to the actual 50% (51.8%).

7.2 Implications for Access Review Challenges

➤ **Decider motivation affects quality (C4):** As described in Section 6.2, the user study participants had a reasonable estimation of their own performance. The user study design is fair, with a planned execution time of 20-30 minutes and no hurdles for *N01: Information Translation* or *N03: Information Visibility*. Still, some deciders submitted results with relatively low quality. The correlations between perceived stress (*FL*, *MD*) and quality (*BA*, *FDR*, *FOR*) may also indicate that decider motivation was an important factor. It must be assumed that poor decider motivation contributes stronger in real-world scenarios with larger scale and poorer information basis, indicating that nudges targeting decider motivation (*N09-N13*) may be a valuable contribution to AR quality.

➤ **Longer reviews are more demanding (C1, C4):** The user study results showed significant correlations between the review duration t and the perceived stress (FL, MD), underlining the importance of a reasonable scale. While the user study already confirms that *N06: Choice Defaults* considerably reduces review time, *N05: Range & Composition* also seems promising. Choice architects should take care not to overwhelm deciders with too many decisions. Distributing review responsibilities to many instead of a few decision-makers might be helpful. Considering *N10: Commitment Facilitation* or splitting reviews into multiple suitable sub-reviews carried out at different times or limiting them to unreviewed or changed authorizations could also improve quality.

➤ **N06: Choice Defaults effectivity does not seem to depend on decision difficulty (C2, C4, C5):** We tried to assess whether the impact of *N06: Choice Defaults* depends on the difficulty of a decision. For this purpose, we grouped the user study decisions by the 160 UPAs and their respective study group (neutral, default accept, default reject), resulting in $3 * 160$ groups of 34 review decisions. We then calculated the error rate and standard deviation for the decisions in the neutral group as indicators of the decision difficulty or uncertainty of UPA. To measure the effect of the default accept nudge for any UPA, we subtract the number of accepts in the neutral group from the amount of accepts in the default accept group. The resulting difference is the amount of *additional* accepts achieved by the nudge. The default reject effectivity was calculated as equivalent to the difference of rejects in the neutral and default reject groups. A Spearman correlation test with a $\alpha = .05$ significance level showed no significant correlation between the indicators for a decision's difficulty and the amount of additional accepts or rejects. The lack of correlation indicates that the effectivity of *N06: Choice Defaults* does not directly depend on the difficulty of a decision.

➤ **Spammers are an error source (C4):** Unlike the user study but in reality, a ground truth of detecting low-quality AR results is not available. Hence, it is helpful to identify "spammers" (deciders actually not trying to achieve an ARP improvement). The user study results suggest two possible ways to determine low-quality AR results: (i) While the review duration t did not correlate significantly with the quality metrics, we found that for the $n = 6$ deciders only taking $t = 6$ minutes or less, the mean BA ($M = 77.1\%$, $SD = 22.1\%$) drops a considerably $\Delta = -14.1\%$ comparing to BA of all participants ($M = 91.2\%$, $SD = 7.9\%$). (ii) Two spammers acted obviously ignorant by blindly accepting or rejecting all authorizations. In real-world scenarios, it might be helpful to use thresholds that, when undercut, classify the review as spam. We do not propose to dismiss such results categorically: it could be correct to accept all authorizations, or a decider could be quick. However, such deciders could be explicitly addressed to improve their result quality, e.g., by applying a custom nudge (like *N13: Empathy Instigation*) or requesting another person to check their decisions.

Table 4: Virtual best and worst advice.

Group	n	R	BA	FDR	FOR
Initial	-	-	.872	.500	.000
Total	102	70.3	.912	.216	.029
Virtual Best Advice	34	71.2	.931	.178	.023
Virtual Worst Advice	34	72.1	.885	.238	.043

Note: n for the participant count and R for the mean of rejected UPAs. BA, FDR and FOR are means for measuring the Access Review Problem (ARP).

➤ **Deciders have the last say (C4):** We re-grouped the user study decisions to simulate reviews with only correct and only incorrect *N06: Choice Defaults* (compare smart defaults [4, 5]). In reality, every decider had to make 160 decisions, of which 80 were TP (should be accepted) and 80 were FP (should be removed). This means that the default accept group had a correct preselection for exactly 80 authorizations, whereas the default reject group had a correct preselection for the other 80 ones. By virtually re-grouping these decisions, we create two sets of $34 * 160$ decisions each, for which one contains only correct default preselections and the other contains only incorrect ones. We then calculated the quality metrics BA, FDR, and FOR for both groups. Unsurprisingly, the virtual best advice group scored a higher overall quality than each of the three real study groups with $BA = 93.1\%$, and the lowest error rates with $FDR = 17.8\%$ and $FOR = 2.3\%$. The virtual worst advice group scored worse than all real groups with $BA = 88.5\%$, $FDR = 23.8\%$, and $FOR = 4.3\%$. However, the virtual best advice group's results are closer to those of all real groups than a perfect result, for which BA would be 100% and both error rates would be 0%. Similarly, the virtual worst advice group did not perform terribly but, in fact, still achieved a mean improvement in the ARP. Results for both groups show that users are affected by the *N06: Choice Defaults* and that the quality of the applied nudge affects the quality of the AR result. However, deciders have the last say and may choose not to follow a default, attenuating the worst assumptions of some interviewed experts. Table 4 summarizes the figures for both virtual groups.

7.3 Two Undesired Responsibility Shifts

Real-world access reviews (without nudge support) assume reflective decision-makers in transparent environments, leading to two assumptions: reflection and transparency [11]. However, the expert interviews and the user study discard both assumptions. For the reflection assumption, experts report several instances of human errors (C4), and the user study shows that deciders are affected by *N06: Choice Defaults*. Additionally, the deciders make errors despite having all the necessary data (even for the best advice in Table 4). For the transparency assumption, experts report the troublesome endeavor to present the information needed (*N01-N03, N09*) as too many or too few details lead to an unclear big picture.

Hansen and Jespersen [19] evaluate ethical considerations for nudge applications by the nudge's transparency and the decider's reflective or automatic mode of thinking. As mentioned earlier, access reviews should strive for transparency and reflective decisions. Access reviews in the real world and those with nudges can fail one of these: the real-world access reviews can lack transparency, and the nudged ones can lack reflective choices. On the one hand, real-world access reviews force reflective decisions as overwhelmed deciders actively need to choose, leading to a lack of transparency and constructing an unpleasant ethical situation. While reflective choices make the deciders fully responsible for their actions, the sheer scale (C1) and frequency (C3) put so many decisions on the table that the actual big picture for the access review becomes non-transparent. Therefore, the deciders have to bear the responsibility for a volume of decisions above their capabilities as human decision-makers, raising ethical concerns. On the other hand, the access reviews with the *N06: Choice Defaults* stay more transparent but allow for less reflective decisions, leading to a responsibility split. As soon as scale (C1) and frequency (C3) make the deciders give up on reflective choices, the choice architect shares responsibility for the decision-makers adopting its defaults.

In summary, neither burdening the deciders with the responsibility of choices they do not comprehend nor splitting the responsibility between the choice architect and the deciders are desired responsibility shifts for access reviews.

7.4 Design Implications for Usability

Following Hansen and Jespersen [19], design implications for future access reviews (with digital nudges) involve facilitating meaningful decisions based on transparency and reflective choices. When applied properly, digital nudges empower deciders to make confident and meaningful decisions with transparent and honest guidance [19]. Most importantly, this implies perceiving access review deciders and their decisions not as hyper-rational but as human, including their strengths and flaws [21, 40]. In the following, we derive three implications for usability based on our results.

➤ **Partition meaningfully:** Several experts find *N05: Range & Composition* relevant as it allows for meaningful partitions of access review decisions. Partitions effectively mitigate the deciders' scale perception and give a context for grouped decisions. Additionally, this allows abstract decisions for the whole partition. For example, deciding to revoke all authorizations of a person can be one meaningful decision instead of rejecting each of its authorizations one by one. Our experts name meaningful ways to partition decisions within access reviews, e.g., people leaving an organization, specific applications, critical authorizations, known past changes, organization-specific attributes, or processes. Ways to determine these partitions can range from choice architects' or deciders' experience to AI-based clustering.

➤ **Apply partition-specific digital nudges:** Digital nudges can be applied individually and combined for each partition. Based on the expert interviews, various digital nudges are suitable. For example, *N06: Choice Defaults* can preselect accepting security-uncritical authorizations (e.g., utility software) or rejecting security-critical ones (e.g., server access). Additionally, security-critical authorizations can be highlighted with a warning by *N02: Information Salience*. Thus, digital nudges can improve each partition's usability to guide access review deciders, also considering individual organizational contexts.

➤ **Query performance perception:** In the user study results, we find a strong correlation in all groups for the objective quality metric *BA* and the deciders' performance self-assessment *PF*. It shows that our user study participants had a reasonable perception of their performance. In contrast, a real-world access review cannot determine *BA* easily, as the underlying ground truth is unknown. This implies querying the deciders' performance self-assessment (*PF*) can be a valid and easy-to-implement estimator for the access review's quality (*BA*).

In summary, transparent digital nudges can guide human decision-makers to make meaningful, confident, and reflective choices. While the positive and negative effects of nudging require careful consideration, their anticipated effects are useful and promising tools for access review designs.

8 Conclusion

In this paper, we investigated digital nudges for access reviews. We formalized the access review problem. Subsequently, we interviewed highly qualified IAM experts to map the expected effects of digital nudges on access review challenges. Furthermore, we conducted a user study with *N06: Choice Defaults*. We found its influence on deciders' behavior in revoking authorizations. Additionally, we achieve time savings (up to 24.3%) and lower frustration. A simple *N06: Choice Defaults* did not significantly influence the overall quality, but it can shift the decisions to more revokes. While these revokes cause some false rejects, false accepts would be worse as they create a false sense of security by legitimating excessive authorizations. For future work, we invite researchers to study the ARP, to investigate other digital nudges of Table 2 or their combinations, or to replicate this study with a larger sample size or smart defaults [4, 5]. In sum, digital nudges are a promising tool to improve access reviews but need careful application.

Availability

For transparency and future research, we make the case study, all collected data, and the analysis of the user study open-source (https://github.com/AccessReview/Availability). In detail, we publish the instructions and data set of the case study, participants' results ($n = 102$), their choices ($n = 16,320$), and the *R* code to replicate our statistical evaluations.

References

[1] William C. Adams. *Conducting Semi-Structured Interviews*, chapter 19, pages 492–505. John Wiley & Sons, Ltd, 2015.

[2] Marvin Auf der Landwehr, Maik Trott, and Christoph von Viebahn. Consumers choice? fostering sustainability in grocery deliveries through digital nudging. In *Twenty-Ninth European Conference on Information Systems (ECIS 2021)*, ECIS 2021, page 1–16. Association for Information Systems, 2021.

[3] Theophilus Azungah. Qualitative research: deductive and inductive approaches to data analysis. *Qualitative Research Journal*, 18(4):383–400, Jan 2018.

[4] Paritosh Bahirat, Yangyang He, Abhilash Menon, and Bart Knijnenburg. A data-driven approach to developing iot privacy-setting interfaces. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces*, IUI '18, page 165–176, New York, NY, USA, 2018. Association for Computing Machinery.

[5] Paritosh Bahirat, Martijn Willemsen, Yangyang He, Qizhang Sun, and Bart Knijnenburg. Overlooking context: How do defaults and framing reduce deliberation in smart home privacy decision-making? In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[6] Basel Committee on Banking Supervision. Basel III: A global regulatory framework for more resilient banks and banking systems, June 2011.

[7] Thomas Baumer, Mathis Müller, and Günther Pernul. System for cross-domain identity management (scim): Survey and enhancement with rbac. *IEEE Access*, 11:86872–86894, 2023.

[8] Michelle Berger, Elias Greinacher, and Linda Wolf. Digital nudging to promote energy conservation behavior - framing and default rule in a smart home app. In *Thirtieth European Conference on Information Systems (ECIS 2022)*, ECIS 2022, page 1–16. Association for Information Systems, 2022.

[9] Kristoffer Bergram, Marija Djokovic, Valéry Bezençon, and Adrian Holzer. The digital landscape of nudging: A systematic literature review of empirical research on digital nudges. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery.

[10] Kelly Caine. Local standards for sample size at chi. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 981–992, New York, NY, USA, 2016. Association for Computing Machinery.

[11] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 23 ways to nudge: A review of technology-mediated nudging in human-computer interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–15, New York, NY, USA, 2019. Association for Computing Machinery.

[12] Federal Financial Supervisory Authority (BaFin). Rundschreiben 05/2023 (BA) - Mindestanforderungen an das Risikomanagement - MaRisk, June 2023.

[13] Sandro Franzoi and Jan vom Brocke. Sustainability by default? nudging carbon offsetting behavior in e-commerce. In *Thirtieth European Conference on Information Systems (ECIS 2022)*, ECIS 2022, page 1–15. Association for Information Systems, 2022.

[14] Ludwig Fuchs and Günther Pernul. HyDRo – hybrid development of roles. In *Information Systems Security*, pages 287–302. Springer Berlin Heidelberg, 2008.

[15] Ludwig Fuchs, Günther Pernul, and Ravi Sandhu. Roles in information security – a survey and classification of the research area. *Computers & Security*, 30(8):748–769, 2011.

[16] Cristina Gena, Pierluigi Grillo, Antonio Lieto, Claudio Mattutino, and Fabiana Vernero. When personalization is not an option: An in-the-wild study on persuasive news recommendation. *Information*, 10(10), 2019.

[17] Abdul Muqeet Ghaffar and Thomas Widjaja. Framing as an app-design measure to nudge users toward infection disclosure in contact-tracing applications. In *Thirty-first European Conference on Information Systems (ECIS 2023)*, ECIS 2023, page 1–16. Association for Information Systems, 2023.

[18] Sebastian Groll, Sascha Kern, Ludwig Fuchs, and Günther Pernul. Monitoring access reviews by crowd labelling. In Simone Fischer-Hübner, Costas Lambrinoudakis, Gabriele Kotsis, A. Min Tjoa, and Ismail

Khalil, editors, *Trust, Privacy and Security in Digital Business*, pages 3–17, Cham, 2021. Springer International Publishing.

[19] Pelle Guldborg Hansen and Andreas Maaløe Jespersen. Nudge and the manipulation of choice: A framework for the responsible use of the nudge approach to behaviour change in public policy. *European Journal of Risk Regulation*, 4(1):3–28, 2013.

[20] Sandra G. Hart. Nasa-task load index (nasa-tlx); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50(9):904–908, 2006.

[21] Jonas Hielscher, Uta Menges, Simon Parkin, Annette Kluge, and M. Angela Sasse. "Employees who Don't accept the time security takes are not aware Enough": The CISO view of Human-Centred security. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 2311–2328, Anaheim, CA, August 2023. USENIX Association.

[22] Linda Hill. How automated access verification can help organizations demonstrate HIPAA compliance: A case study. *J Healthc Inf Manag*, 20(2):116–122, 2006.

[23] Vincent C. Hu, David Ferraiolo, Rick Kuhn, Arthur R. Friedman, Alan J. Lang, Margaret M. Cogdell, Adam Schnitzer, Kenneth Sandlin, Robert Miller, Karen Scarfone, et al. Guide to attribute based access control (abac) definition and considerations (draft). Technical report, National Institute of Standards and Technology, 2014.

[24] Dennis Hummel and Alexander Maedche. How effective is nudging? a quantitative review on the effect sizes and limits of empirical nudging studies. *Journal of Behavioral and Experimental Economics*, 80:47–58, 2019.

[25] Matthias Hummer, Sebastian Groll, Michael Kunz, Ludwig Fuchs, and Günther Pernul. Measuring identity and access management performance - an expert survey on possible performance indicators. In *Proceedings of the 4th International Conference on Information Systems Security and Privacy*, pages 233–240. SCITEPRESS - Science and Technology Publications, 2018.

[26] Pooya Jaferian, Hootan Rashtian, and Konstantin Beznosov. To authorize or not authorize: Helping users review access policies in organizations. In *Proceedings of the Tenth USENIX Conference on Usable Privacy and Security*, SOUPS '14, page 301–320, USA, 2014. USENIX Association.

[27] Mathias Jesse and Dietmar Jannach. Digital nudging with recommender systems: Survey and future directions. *Computers in Human Behavior Reports*, 3:100052, 2021.

[28] Eric J. Johnson, Suzanne B. Shu, Benedict G. C. Dellaert, Craig Fox, Daniel G. Goldstein, Gerald Häubl, Richard P. Larrick, John W. Payne, Ellen Peters, David Schkade, Brian Wansink, and Elke U. Weber. Beyond nudges: Tools of a choice architecture. *Marketing Letters*, 23(2):487–504, Jun 2012.

[29] Shelia M. Kennison, Ian T. Jones, Victoria H. Spooner, and D. Eric Chan-Tin. Who creates strong passwords when nudging fails. *Computers in Human Behavior Reports*, 4:100132, 2021.

[30] Sascha Kern, Thomas Baumer, Ludwig Fuchs, and Günther Pernul. Maintain high-quality access control policies: An academic and practice-driven approach. In Vijayalakshmi Atluri and Anna Lisa Ferrara, editors, *Data and Applications Security and Privacy XXXVII*, pages 223–242, Cham, 2023. Springer Nature Switzerland.

[31] Sascha Kern, Thomas Baumer, Sebastian Groll, Ludwig Fuchs, and Günther Pernul. Optimization of access control policies. *Journal of Information Security and Applications*, 70:103301, 2022.

[32] Stefan Meier, Ludwig Fuchs, and Günther Pernul. Managing the access grid - a process view to minimize insider misuse risks. In *11th International Conference on Wirtschaftsinformatik (WI2013)*, pages 1051–1065, 2013.

[33] Christian Meske and Tobias Potthoff. The dinu-model - a process model for the design of nudges. In *European Conference on Information Systems*, pages 2587–2597, 06 2017.

[34] Tobias Mirsch, Christiane Lehrer, and Reinhard Jung. Making digital nudging applicable: The digital nudge design method. In *International Conference on Information Systems*. AIS, 2018.

[35] Robert Münscher, Max Vetter, and Thomas Scheuerle. A review and taxonomy of choice architecture techniques. *Journal of Behavioral Decision Making*, 29(5):511–524, August 2015.

[36] OWASP Top 10 team. Owasp top10, 2021. Accessed: 11/15/23.

[37] Simon Parkinson and Saad Khan. A survey on empirical security analysis of access-control systems: A real-world perspective. *ACM Comput. Surv.*, 55(6), dec 2022.

[38] Alexander Puchta, Fabian Böhm, and Günther Pernul. Contributing to current challenges in identity and access management with visual analytics. In Simon N. Foley, editor, *Data and Applications Security and Privacy*

XXXIII, pages 221–239, Cham, 2019. Springer International Publishing.

[39] Alexander Puchta, Sebastian Groll, and Günther Pernul. Leveraging dynamic information for identity and access management: An extension of current enterprise iam architecture. In *Proceedings of the 7th International Conference on Information Systems Security and Privacy - ICISSP*, pages 611–618, Online Streaming, 2021. INSTICC, SciTePress.

[40] Ita Ryan, Utz Roedig, and Klaas-Jan Stol. Unhelpful assumptions in software security research. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, CCS '23, page 3460–3474, New York, NY, USA, 2023. Association for Computing Machinery.

[41] Pierangela Samarati and Sabrina Capitani de Vimercati. Access control: Policies, models, and mechanisms. In Riccardo Focardi and Roberto Gorrieri, editors, *Foundations of Security Analysis and Design*, pages 137–196, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg.

[42] William Samuelson and Richard Zeckhauser. Status quo bias in decision making. *Journal of risk and uncertainty*, 1:7–59, 1988.

[43] Ravi S. Sandhu. Role-based access control. portions of this chapter have been published earlier in sandhu et al. (1996), sandhu (1996), sandhu and bhamidipati (1997), sandhu et al. (1997) and sandhu and feinstein (1994). In Marvin V. Zelkowitz, editor, *Advances in Computers*, volume 46 of *Advances in Computers*, pages 237–286. Elsevier, online, 1998.

[44] Armando Schär and Katarina Stanoevska-Slabeva. Application of digital nudging in customer journeys - A systematic literature review. In *25th Americas Conference on Information Systems, AMCIS 2019, Cancún, Mexico, August 15-17, 2019*. Association for Information Systems, 2019.

[45] Christoph Schneider, Markus Weinmann, and Jan vom Brocke. Digital nudging: Guiding online user choices through interface design. *Commun. ACM*, 61(7):67–73, jun 2018.

[46] Daniel Servos and Sylvia L. Osborn. Current research and open problems in attribute-based access control. *ACM Comput. Surv.*, 49(4), jan 2017.

[47] Bingyu Shen, Tianyi Shan, and Yuanyuan Zhou. Improving logging to reduce permission Over-Granting mistakes. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 409–426, Anaheim, CA, August 2023. USENIX Association.

[48] Cass R. Sunstein. The council of psychological advisers. *Annual Review of Psychology*, 67(1):713–737, 2016. PMID: 26393867.

[49] Barnabas Szaszi, Anna Palinkas, Bence Palfi, Aba Szollosi, and Balazs Aczel. A systematic scoping review of the choice architecture movement: Toward understanding when and why nudges work. *Journal of Behavioral Decision Making*, 31(3):355–366, 2018.

[50] Richard H. Thaler and Cass R. Sunstein. *Nudge: Improving decisions about health, wealth, and happiness.* Nudge: Improving decisions about health, wealth, and happiness. Yale University Press, New Haven, CT, US, 2008.

[51] United States Congress. Health Insurance Portability and Accountability Act of 1996, 1996.

[52] United States Congress. Sarbanes-Oxley Act of 2002. Corporate responsibility, 2002.

[53] Markus Weinmann, Christoph Schneider, and Jan vom Brocke. Digital nudging. *Business & Information Systems Engineering*, 58(6):433–436, Dec 2016.

[54] Tianyin Xu, Han Min Naing, Le Lu, and Yuanyuan Zhou. How do system administrators resolve access-denied issues in the real world? In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 348–361, 2017.

[55] Sarah Y. Zheng and Ingolf Becker. Checking, nudging or scoring? evaluating e-mail user security tools. In *Nineteenth Symposium on Usable Privacy and Security (SOUPS 2023)*, pages 57–76, Anaheim, CA, August 2023. USENIX Association.

[56] Samira Zibaei, Dinah Rinoa Malapaya, Benjamin Mercier, Amirali Salehi-Abari, and Julie Thorpe. Do password managers nudge secure (random) passwords? In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 581–597, Boston, MA, August 2022. USENIX Association.

[57] Samira Zibaei, Amirali Salehi-Abari, and Julie Thorpe. Dissecting nudges in password managers: Simple defaults are powerful. In *Nineteenth Symposium on Usable Privacy and Security (SOUPS 2023)*, pages 211–225, Anaheim, CA, August 2023. USENIX Association.

Appendix

A Expert Interviews

A.1 Interview Script

I. Intro Section

Interview partner

- *What is your job position at organization XY?*
- *What is your IAM experience (years, clients, access review projects, managed identities, etc.)?*

Access review and its problems

- *Estimate the ratio of excessive granted access.*
- *Name 2-3 major challenges for access reviews.*

II. Explanation Section

- *Explain to the participant the access review challenges of Jaferian et al. [26]. Connect them to the major challenges of access review the participant named before.*
- *Explain to the participant digital nudges in general.*

III. Workshop Section

Mapping digital nudges and access review challenges
For each nudge in Table 5

1. *Explain the nudge and give an example fitting for the interview participant's environment.*
2. *The participant then freely reflects on the digital nudge and their relationship on access review challenges.*
3. *Finally, the participant rates each access review challenge, anticipating a very positive (+2), positive (+1), neutral (0), negative (-1), or very negative (-2) effect.*

Table 5: Digital Nudges [27] presented to the experts for mapping them to access review challenges [26].

Nudges	C1	C2	C3	C4	C5
Decision Information					
N01: Information Translation					
N02: Information Salience					
N03: Information Visibility					
N04: Information Phrasing					
Decision Structure					
N05: Range & Composition					
N06: Choice Defaults					
N07: Option Consequences					
N08: Option-related Effort					
Decision Assistance					
N09: Reminders					
N10: Commitment Facilitation					
Social Decision Appeal					
N11: Messenger Reputation					
N12: Social Reference Point					
N13: Empathy Instigation					

Wrap-up

- *Name your TOP 3 digital nudges benefiting access review challenges.*

A.2 Codebook

We apply deductive and inductive coding to the expert interviews. The feasibility of digital nudges (based on the collection of Jesse and Jannach [27]) for access reviews suffice as interview questions. The access review challenges of Jaferian et al. [26] suffice as deductive codes, which we applied a priori to the interviews. Therefore, we trained and asked the interview partners about these challenges and asked for a Likert scale-based rating (2 (best), 1, 0, -1, -2 (worst)). The experts answered with different arguments, for which we extracted inductive codes. The rating for digital nudge, challenge and the inductive codes are detailed in the codebook (Table 6).

B User Study

B.1 Data Set

For the user study, we used a crafted data set (160 UPAs). We can pinpoint which UPAs are correctly (TP) and incorrectly (FP) assigned. Figure 5 (using a grid representation based on [32]) depict the data set. A processable format is available at GitHub.

B.2 Ground Truth Document

Access Review Case Study
You work as a busy head of the marketing department in a large industry company with many concurrent projects to maximize the income for your company. Your time is limited, and you have marketing goals to fulfill.
The security teams reminded you via email that your company is legally required (compliance) to review the permission assignments for the employees in your department. You must follow the **principle of least privilege**: Employees must have permissions required for their job, but not more. If you decide to revoke an excessive permission for one of your employees, the employee will no longer be able to access the associated resources by tomorrow.
While the security team points out that any excessive permission poses a security threat, you are aware that missing ones might prevent your employees from working until they re-obtain it via a time-consuming help-desk or self-service request.
The marketing department consists of three teams:

I. Graphic design team

- Create and edit images for the company's media and advertisement presence. This includes banners, logos, websites, or campaign designs that are used in advertisements or social media posts.
- Require a Photoshop license to work.

II. Social media team
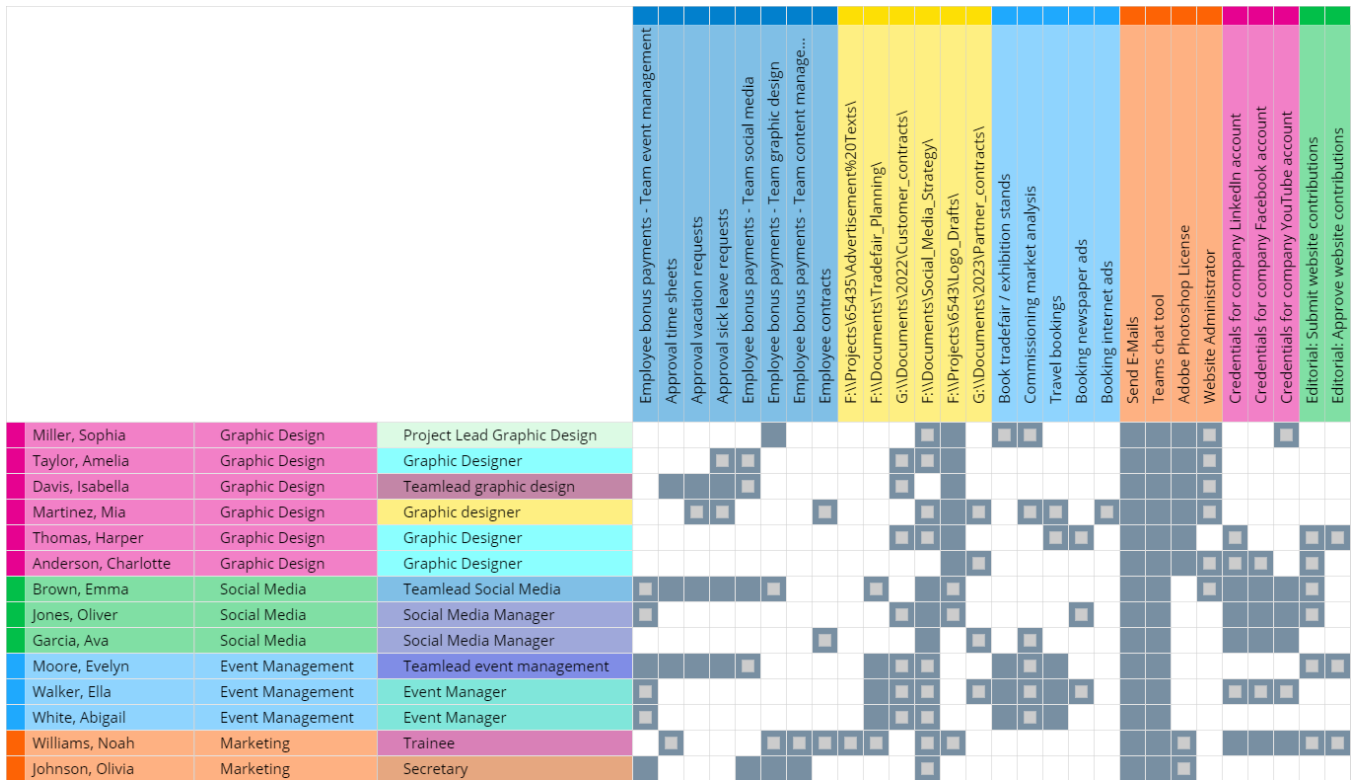
Figure 5: Grid visualization [32] of the user study data set. Blue cells resemble TP, gray ones FP, white ones TN, and FN were not present in the data set.

- Manage the company's social media accounts.

- Need to communicate with potential customers, candidates for recruiting, and partners online.

III. Event management team

- Organize trade fairs and partner events across West and Central Europe.

- Book trade fair stands.

- High self-organization; often need to attend remote events without long preparation.

IV. Department hierarchies

- Every team is led by a team lead who overlooks the employee's attendance and work results.

- Team leads have an annual budget for bonus payments, which they can distribute among their team members based on last year's performance. The secretary reads the specified bonus payments defined by the team leads from the HR system and arranges for the salary to be posted.

- The department's trainee used to intern in the graphic design team. Now, he is working in the social media team.

V. Misc

- Everybody communicates with MS Teams and Outlook.

- You can sort the columns.

B.3 Screenshots Access Review

We used three configurations of the access reviews with the same data basis. Figure 6, the neutral default, has two white buttons without a preselection. Figure 7 displays the default accept with a preselected *Approve*. Figure 8 shows the default reject with a preselected *Remove*.

B.4 Statistical Analysis

Figure 9 depicts a pair plot for each metric separated for their group. Green shows the default accept group, red the default reject, and blue the neutral one. The upper right part depicts Spearman correlations. The stars indicate the significance levels as "***": $p < .001$; "**": $p < .01$; "*": $p < .05$, and "." $p < .1$. The lower left depicts local regressions. Finally, the diagonal, the first row and column show metric distributions.

Figure 6: Screenshot of the neutral group for the user study.

Figure 7: Screenshot for the accept group.

Figure 8: Screenshot for the reject group.

Table 6: Codebook for expert interviews.

N	C	Likert	Inductive Codes
N01	C1	1	Understandability (E02, E03, E04, E09); No effect (E03, E06, E08, E10); Feel-Good (E02, E03); Uniqueness (E04, E09); Structure (E09)
N01	C2	2	Understandability (E02, E05, E06, E07, E08, E09, E10); Mental Load (E03, E05, E07, E09); Acceptance (E05, E07, E10); Wording (E05, E06, E07)
N01	C3	1	Recognition (E01, E04, E05, E06, E09); Learning (E04, E05, E09); Feel-Good (E05, E06)
N01	C4	2	Understandability (E02, E05, E07, E08, E09, E10)
N01	C5	0	Understandability (E05, E06); Recognition (E04)
N02	C1	1	Focus (E01, E02, E04, E05, E09); No effect (E06, E08, E10)
N02	C2	0	Focus (E06, E09, E10); No effect (E03, E07, E09)
N02	C3	1	Economic Efficiency (E01, E02, E04, E07, E09); Focus (E01, E02); Acceptance (E09)
N02	C4	1	Focus (E03, E05, E06, E07, E08, E09, E10); Algorithm-Quality (E03, E06, E09); Backlash (E03, E06, E09)
N02	C5	2	Focus (E03, E04, E05, E07, E10); Algorithm-Quality (E09)
N03	C1	1	More relevancy (E03, E04, E05, E07, E09); Less confusion (E03, E04, E05, E09); No reduction of decisions (E06, E08)
N03	C2	2	Showing more data (E01, E05, E08, E09, E10); Relevancy (E03, E04, E06)
N03	C3	0	Run-time (E05); Recognition (E07)
N03	C4	1	Mistake mitigation (E05, E07, E09, E10); Focus (E06, E07)
N03	C5	2	Showing more data (E01, E07, E09, E10); Need to know (E07, E10)
N04	C1	0	Insecurities of decision-maker (E09); Sense of responsibility (E07)
N04	C2	-1	Context-Awareness (E05, E07, E08, E09, E10); Bias (E04, E05, E09), Base direction (E05, E09)
N04	C3	0	Acceptance (E07)
N04	C4	1	Acceptance (E06; E08; E09; E10); Focus (E06, E09, E10); Pressure (E02)
N04	C5	0	Focus (E06, E07, E10)
N05	C1	2	Similarities (E01, E03, E04, E05, E07, E08, E09); Overhead (E08, E10)
N05	C2	1	Focus (E04, E08, E09, E10); Audience (E08, E09)
N05	C3	1	Economic Efficiency (E03, E05, E06); More Tasks (E09, E10)
N05	C4	2	Focus (E01, E03, E05, E06, E07, E09, E10); Similarities (E01, E05, E06, E07); Smaller Batches (E09, E10)
N05	C5	2	Exceptional Case Detection and View (E02, E03, E04, E07, E09)
N06	C1	2	Less work (E01, E02, E04, E05, E06, E09); No reduction of decisions (E07, E10)
N06	C2	-2	Recommendation Fallacy (E02, E04, E05, E06, E07, E09, E10); Recommendation Support (E06, E09)
N06	C3	2	Less work (E01, E02, E04, E05, E06, E08, E09)
N06	C4	-2	Less diligence/Focus (E01, E02, E04, E05, E06, E07, E09, E10); Recommendation Fallacy (E02, E04, E05, E06, E07, E09, E10)
N06	C5	0	Not in Focus (E05, E07, E10); Default handling (E06, E09); Special treatment (E04)
N06	Misc		Not Compliant (E01, E03, E07, E09); Needs good recommendation (E01, E03, E08, E09); Is it really a decision? (E03, E07, E09)
N07	C1	0	Speed (E01, E09)
N07	C2	-1	Recommendation Fallacy (E09)
N07	C3	1	Speed (E01, E04, E09); Gamification (E04, E05, E09); Feel-Good (E04, E07); Acclimatation (E09)
N07	C4	-1	Pressure (E01, E03, E07, E09, E10); Recommendation Fallacy (E06, E07, E09, E10); Less diligence (E01, E03, E07)
N07	C5	-1	Recommendation Fallacy (E07, E09); Fairness for disadvantaged individuals (E09)
N08	C1	-1 / 1	Ambivalence (E01, E03, E05, E06, E07, E08, E09, E10); Economic Efficiency (E02, E03, E05, E07, E08); Acceptance (E04, E07, E09, E10)
N08	C2	1 / -1	Ambivalence (E01, E03, E05, E06, E07, E08, E09, E10)
N08	C3	-1 / 1	Ambivalence (E01, E03, E05, E06, E07, E08, E09, E10); Economic Efficiency (E02, E03, E05, E07, E08); Acceptance (E04, E07, E09, E10)
N08	C4	1 / -1	Ambivalence (E01, E03, E05, E06, E07, E08, E09, E10); Economic Efficiency (E02, E03, E05, E07, E08); Acceptance (E04, E07, E09, E10)
N08	C5	1 / -1	Ambivalence (E01, E03, E05, E06, E07, E08, E09, E10)
N09	C1	0	No effect (E03, E07); More participation (E04, E10);
N09	C2	1	Instructions and Goals (E02, E03, E04, E05, E07, E08, E09, E10); Spam (E01, E02, E03, E07, E09, E10);
N09	C3	2	Spam (E01, E02, E03, E07, E09, E10); Attention (E04, E06, E07, E09, E10)
N09	C4	-1	Revisit (E03, E07, E08); Pressure (E05); Multi-Channel (E07, E09, E10); Audience (E03, E09)
N09	C5	0	Open Task (E01); No effect (E07)
N10	C1	1	Combination with N05 - Commitment for partitions (E02, E04, E06, E07, E08, E09, E10)
N10	C2	0	Autonomic planning and understanding (E04, E05, E08, E09)
N10	C3	1	Combination with N05 - Sub-Deadlines for partitions (E05, E07, E08, E09, E10); Comfort (E02, E06, E07, E08, E09, E10)
N10	C4	1	Focus (E04, E06, E07, P8, E10); Comfort (E02, E06, E07, E08, E09, E10)
N10	C5	0	Focus (E07, E10)
N11	C1	1	Endurance (E01, E02, E03, E04, E05, E07, E09, E10); Trust (E02, E03, E07, E08, E09)
N11	C2	2	Approachable IAM team (E01, E02, E03, E04, E05, E07, E08, E09, E10)
N11	C3	1	Endurance (E01, E02, E03, E04, E05, E07, E09, E10); Trust (E02, E03, E07, E08, E09)
N11	C4	2	Approachable IAM team (E01, E02, E03, E04, E05, E07, E08, E09, E10); Acceptance (E01, E02, E03, E04, E08, E10)
N11	C5	2	Approachable IAM team (E01, E02, E03, E04, E05, E07, E08, E09, E10)
N12	C1	0	Endurance (E02, E03, E06, E09); Backlash (E09)
N12	C2	2	Approachable Peer-Group (E02, E03, E04, E07, E08, E09, E10)
N12	C3	0	Endurance (E02, E03, E06, E09); Backlash (E09)
N12	C4	1	Acceptance (E02, E03, E06, E07, E09, E10); Peer-Pressure (E03, E10)
N12	C5	2	Approachable Peer-Group (E02, E03, E04, E07, E08, E09, E10)
N12	Misc		Similarity to N11 messenger reputation (E01, E05)
N13	C1	1	Feel-Good (E02, E04, E06, E07, E09)
N13	C2	1	Feedback on odd behavior (E02, E03, E04, E07, E09)
N13	C3	1	Feel-Good (E02, E04, E06, E07, E09)
N13	C4	1	Feel-Good (E02, E05, E07, E08); Focus (E02, E04, E07, E09)
N13	C5	0	Feel-Good (E05)
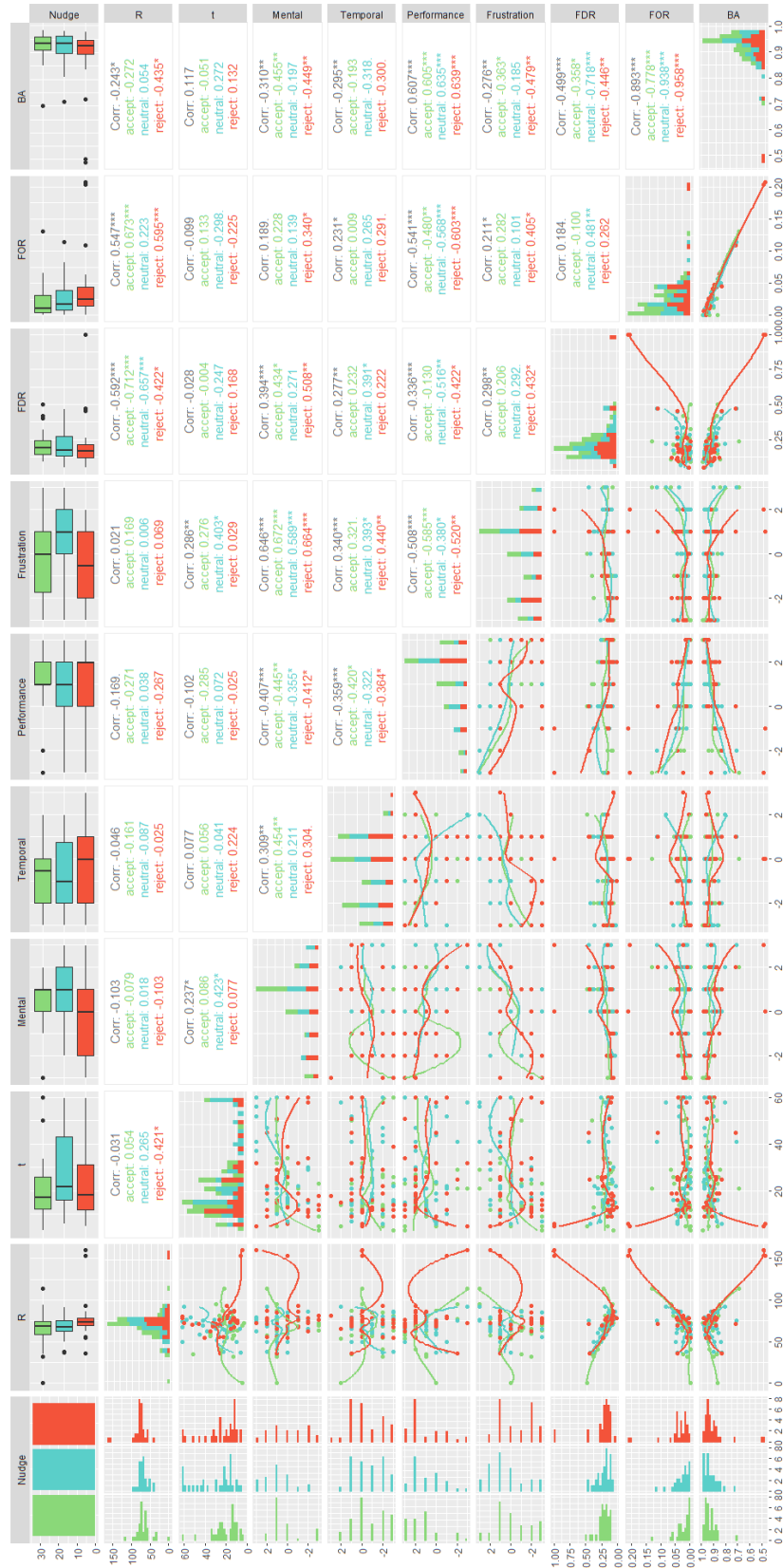
Figure 9: Pair plot of correlations (Spearman) and local regressions for the user study.

Can Johnny be a whistleblower?
A qualitative user study of a social authentication Signal extension in an adversarial scenario

Maximilian Häring
University of Bonn

Julia Angelika Grohs
University of Bonn

Eva Tiefenau
Fraunhofer FKIE

Matthew Smith
University of Bonn, Fraunhofer FKIE

Christian Tiefenau
University of Bonn

Abstract

To achieve a higher level of protection against person-in-the-middle attacks when using common chat apps with end-to-end encryption, each chat partner can verify the other party's key material via an out-of-band channel. This procedure of verifying the key material is called an authentication ceremony (AC) and can consist of, e.g., comparing textual representations, scanning QR codes, or using third party social accounts. In the latter, a user can establish trust by proving that they have access to a particular social media account. A study has shown that such social authentication's usability can be very good; however, the study focused exclusively on secure cases, i.e., the authentication ceremonies were never attacked. To evaluate whether social authentication remains usable and secure when attacked, we implemented an interface for a recently published social authentication protocol called SOAP. We developed a study design to compare authentication ceremonies, conducted a qualitative user study with an attack scenario, and compared social authentication to textual and QR code authentication ceremonies. The participants took on the role of whistleblowers and were tasked with verifying the identities of journalists. In a pilot study, three out of nine participants were caught by the government due to SOAP, but with an improved interface, this number was reduced to one out of 18 participants. Our results indicate that social authentication can lead to more secure behavior compared to more traditional authentication ceremonies and that the scenario motivated participants to reason about their decisions.

1 Introduction

End-to-end encryption (E2EE) is a well-known and broadly applied technology in messaging apps. Its implementation helps to improve the privacy of billions of people. However, E2EE cannot provide authenticity without the interaction of users. To have authenticity, chat partners must ensure that the correct key material is used, i.e., the service provider is not tampering with the keys to mount a person-in-the-middle (PITM) attack.

The task of comparing the key material of the communication partners, e.g., by meeting in person and showing them, is called an *authentication ceremony (AC)*. By correctly carrying out an AC, users can be sure that they are talking confidentially with the right person. However, the default in current messaging apps is to trust the first keys given to users by the provider without encouraging an AC [2] and inform users when these keys change. Studies show that few users run authentication ceremonies, and many users do not know the cryptographic notion of authentication and how to handle the corresponding ceremonies [3, 5].

A possible reason why few users have a reason to verify keys is that even without verification E2EE provides a good level of protection as mass surveillance is resource-hungry and disincentivized for the attacker; getting caught is fairly likely due to key-change notifications that can be noticed by the provider or experts, e.g., facilitated by key transparency [9, 13, 28]. However, targeted surveillance can still be a threat as it is technologically possible, and the risk-benefit ratio for the attacker could be worthwhile. Consequently, we believe that if there is a need for authentication ceremonies, it is most pressing in high-risk scenarios, e.g., when one is a political dissident, a whistleblower, or a government employee. While the single tasks that are necessary for ACs can be done quickly and with rather low false-acceptance rates [18, 25], studies provide evidence that current authentication ceremonies are difficult and error prone [5, 6, 17, 27].

A fairly new solution for remote[1] authentication, "social

---

[1]"Remote" refers to a setting where the two communication partners carry

authentication (SA)" was suggested by prior research and leverages social networking sites as a trust anchor [8, 11, 24].

The idea behind this solution is to reduce the verification task to something users can already do and intuitively grasp. For SA, users do not need to compare key material directly; instead, they must decide which identity provider, e.g., a social media site, to trust and recognize an already known account. As such, users need to have knowledge about the contact they want to authenticate and know their identifier (e.g., Alice42) on the chosen identity provider (e.g., facebook.com). Vaziripour et al. [24] tested the concept in a laboratory study and found the approach to have good usability. They reported that participants found the concept convenient and matched "how participants thought of verification." However, their solution was tested under ideal conditions, i.e., without any attackers. Nevertheless, the researchers noted that SA makes identity spoofing and impersonation attacks possible. Currently, no work on SA in an attack scenario exists. To fill this knowledge gap, we conducted a user lab study where we simulated an attack scenario and compared SA to the already established ACs of key fingerprint comparison and QR codes.

This work contributes a novel methodology for comparing ACs and an interface to make social authentication similarly usable as safety numbers or QR codes. We extend the existing literature on ACs and how they are researched by testing an attack scenario in a **user study** of a SA approach. We were especially interested in the participants' reactions toward impersonation attacks, i.e., how often they would notice the attack and how they would proceed with a given task.

We created a scenario that resembles, more closely than previous work, a realistic use case for users needing an authentication method. To motivate the participants to authenticate and mimic real-world situations, they had to act as whistleblowers in an authoritarian regime and contact journalists. This **study design with a scenario with reasonable participant motivation** allowed us to observe the entire process of the authentication ceremonies. In contrast to Vaziripour et al.'s study [24], which proposed a form of SA, Linker et al. [11] formally defined SA and presented a protocol with proven security properties. They called the protocol SOAP and developed a prototype that worked, with limitations, within the current internet eco-system. This means our results can be directly applied to their prototype and hopefully increase the security of users.

During our analysis, we were guided by the following research questions:

RQ1 - Detection: How resistant is SOAP to impersonation attacks?

RQ2 - Reaction: How do participants react to a detected impersonation attack?

RQ3 - Perception: What are users' perceptions of SOAP (usability, trustworthiness), with a focus on identity providers?

---

out an AC without meeting in person. Although we phrase ACs as a task for two users, it often works similarly for more than two.

In a pilot study, nine participants used a simple interface based on the protocol proposed by Linker et al. [11], which was implemented as an extension to the Signal app [19]. Many participants failed to use SA correctly when under attack. After analyzing the results, we improved the interface and recruited 18 participants. Although our design improved the results so that only one participant behaved insecurely because of SOAP, six of the lab study's 18 participants failed to detect a PITM for other reasons. If applied to the real world, this would mean that they would be in danger if they were to rely on a tool like Signal for confidentiality.

The rest of this paper is structured as follows: In Section 2, we provide a short overview of relevant authentication methods, their shortcomings, and the concept of SA. In Section 3, we present the user study, and in Section 4, we discuss implications and further directions for research and messaging app developers.

2 Related Work and Background

In this section, we summarize ACs in the messaging app domain and related work about them to put social authentication into context.

2.1 Authentication Ceremonies

Comparing key material, a process called authentication ceremony (AC), has scarcely changed in the last few years. Via an AC, a PITM attack can be detected, e.g., if the attacker uses a key substitution attack [6].

Material for comparison is always based on the public key, but the visualization differs among apps [2, 6]: Signal initially displayed two public key fingerprints before changing to concatenation and currently displays a single safety number [14]. In addition to that, Signal also offers a QR code, which is a different representation of the single safety number. A recent version of Telegram (iOS 10.3.1) shows a scannable identicon, similar to a QR code, and a hex notation "generated from hashes of the DH secret chat keys" [22]. During phone calls, emojis are shown [21].

The success of an AC has its challenges. Herzberg et al. [6] structured these as deciding that a ceremony is needed, finding the ceremony in the user interface, executing the ceremony, understanding the result, and acting on it.

As it is assumed and evidenced [17, 26] that users struggle with ACs, studies looked at each of the steps in the process and tried to improve them. Vaziripour et al. [25] worked on guiding users to the ceremony interface. With opinionated design, they were able to lead 90% of their study participants to the ceremony. Wu et al. [27] worked on users' comprehension of safety number change notifications and found a need to communicate the possible risk to users as a motivation and basis to decide. Shirvanian et al. [18] studied whether the comparison act itself could be a problem. They found

evidence that in a remote setting (i.e., when users do not sit next to each other), comparison can be an error-prone task, mainly because users need to compare codes between two apps on the same device, with the need to remember the code. Tan et al. [20] and Livsey et al. [12] researched how different visualizations impact the comparison act. Although Livsey et al. found that their participants did not make many mistakes, Tan et al. found that visualization can greatly impact the outcome in an attack scenario, with success rates for the attacker varying between 6% and 72%. All these studies and the methods used rely on the same AC principle: a direct manual exchange and comparison of key material to authenticate the communication partner. As described in the next section, social authentication relies on a different principle.

2.1.1 Social Authentication

In the literature, two different topics are referred to as social authentication. According to Jain et al. [7], SA describes when Alice wants to log in to a service and another user, Bob, who is connected to Alice on the (social media) platform, is asked whether they are allowed to. This can be triggered, e.g., as a step in a risk-based authentication scheme. However, Vaziripour et al. [24] described SA as an AC completed through "social media." In SA, public key material is distributed through a social media provider. In this paper, we refer to this second notion of SA as an AC.

Concept   With this AC, the challenge of the ceremony is shifted from selecting a secure channel, exchanging the key material, and comparing the fingerprints to deciding what provider to trust and recognizing an identifier.

An early application where this notion of SA is in place is Keybase. On Keybase, a user can provide proof of having access to an account by posting material on it. Afterward, other Keybase users can decide whether proof of access to that account is enough for them to identify the person [8]. Vaziripour et al.'s [24] proposed system is very similar. The researchers envisioned that Signal users would log in to their social media accounts during configuration, and the public key material would be posted there. Similar to the scheme utilized by Keybase, this would allow observers to see the material. For example, if Alice wants to check whether the E2EE on Signal is PITM-free and authentic, they could check whether Bob has provided a reference account on a trusted social media platform, in the following called identity provider (IdP). As the key material is posted online, it can be compared automatically and asynchronously. The decision Alice has to make is whether they trust the IdP and whether the account provided by Bob belongs to the person they want to contact.

Studies   Vaziripour et al. [24] tested their idea in a lab study (21 participant pairs) and an online survey (N=421). They let the participants communicate via Signal and, if needed,

guided them to the AC. Here, the participants were able to choose between three verification methods: *social media* (social authentication), *in person*, and *phone call*. The participants were allowed to choose from all three methods and were asked to use the remaining two after selecting one. The researchers found that the *social media* verification method had the best Single-Ease-Question (SEQ) score but was less trusted than the *in person* and *phone call* methods. Additionally, the participants chose the *in person* method first (n=20) more often than the *social media* method (n=12). The average configuration time of the *social media* method was 2 minutes and 32 seconds. On average, verification (which, in this case, meant looking at profile names and pictures) took 34 seconds. Varziripour et al. concluded that social media was not perceived as a highly trustworthy provider of authentication, but the participants liked the asynchronicity, that it worked remotely, and that it was partially automated. As the challenge of the AC changes, so does the attack surface. The participants in Varziripour et al.'s study mentioned the attack vector of fake profiles, which indeed seems to be a major challenge for SA. Additionally, the key material has to be public. This could be problematic for some users due to privacy considerations.

SOAP   Another recent proposal, "SOAP" [11], mitigates the need to have the key material public and aims to find a way to bootstrap SA in the current internet without too much effort from the provider's site. Hence, SOAP utilizes IdPs, not necessarily social media providers. An identity provider (IdP) could be any entity providing an OpenID Connect service, hoping for a relatively fast and easy adoption. If Alice wants to check the security of the chat with Bob, Alice asks Bob to prove that they have control over an account at a specific (listed) IdP. This can be done by just sending a message to Bob to do so or utilizing a UI flow as proposed in this paper. After that, Bob's client asks an IdP to sign the chat's safety number in combination with his account on the platform. Bob has to first log in to the provider before the IdP signs a request. The signing is done automatically. Bob then forwards the signed message returned from the IdP to Alice. Afterward, Alice has a statement from the IdP that says: With whom you are talking to, identified by this safety number, has control over account "XYZ" on my platform. More technically, Bob's client starts an OpenID Connect authorization code flow with the salted hash of the safety number and a nonce. The resulting ID token (including the nonce) and the salt are forwarded to Alice. Alice's client can now check whether the safety number it has matches the one incorporated within the token. Alice then has to decide whether the identity provided is as expected and wanted. To the best of our knowledge, currently, no research on SA attack scenarios exists. We fill this gap in the remainder of this paper.
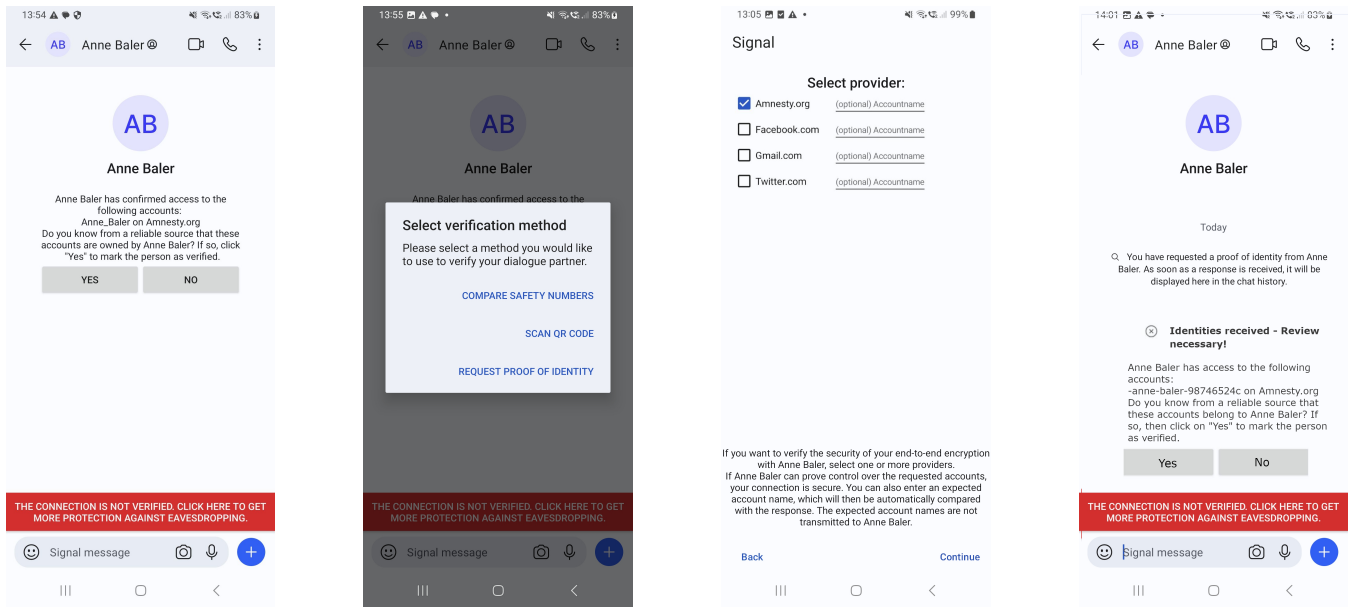
(a) New chat for participants in the **pre-registration** condition. They saw a non-requested SOAP answer. Otherwise, the chat was empty. The **red banner** nudged the participant to find the ceremonies.

(b) Menu that opens if participant clicks the red banner. The first two options led to the currently implemented safety number site with slightly modified text. The third option opens a SOAP request interface.

(c) A SOAP request can be made by selecting an IdP and, if wanted, adding an expected identifier. By continuing, a SOAP request is sent. If an identifier is added, the recipient cannot see this.

(d) If it cannot be determined automatically whether the identifier is correct, the user must decide.

Figure 1: Translated screenshots of the UI used for the lab study.

3 User Study

We tested SOAP in a lab study with a preceding pilot study. We implemented and started with a simple SOAP [11] interface for the Android Signal app. Based on a pilot study (n = 9), we adapted the interface. At last, we ran a lab study with 18 participants. This section describes the resulting SOAP interface and the study design and presents results from the pilot and the lab study.

3.1 Technical Implementation - UI

Linker et al. [11] presented with the protocol an accompanying prototype that implements the technical protocol but does not hint at its capabilities to the user. Only one button in the app's share menu suggested the existence of SA. So, the verifier has to know that SOAP exists and somehow agree with the to-be-verified person what identities and IdPs are available and then ask for proof. For this study, we were not interested in whether people could find the icon, and we did not want to explain the idea in a workshop. Seeing that the prototype's design was not ready for our purposes, we adapted it to the needs of our study. We used the Signal app because the prototype builds on it, it is open source, and previous

studies also used Signal. In the following sections, we detail relevant elements of the technical implementation from after the pilot study. An earlier version of the interface can be seen in the Appendix (Figure 3).

3.1.1 Hint to the Ceremonies

To test SA, we wanted to point the participants directly to the relevant parts of the interface. Vaziripour et al. [25] successfully led users to the ACs with a clear, visible red banner above the text entry field in chat views, and we adopted the same method (see Figure 1a). A click on the banner triggered a dialog with the three ceremonies (see Figure 1b): Safety Number, QR-Code, and SOAP.

3.1.2 QR Code and Safety Number

When the participants clicked on the QR code or safety number button, they landed on the slightly modified safety number page in the chat settings as in recent versions of Signal (see Figure 5d in the Appendix), where a QR code could be scanned, and the chat's safety number read. A message must have been exchanged with the contact for a chat to have a safety number. If the participants tried to access this page

without prior communication, a popup reminded them that a first exchange must happen. Please note that although Signal provides a unique safety number per chat, it is only a concatenation of two per-user numbers. So, just the half belonging to the contact has to be checked. We added an explanation of this to Signal's settings page.

3.1.3 SOAP - The Social Authentication Protocol

Choosing "request proof of identity" opened a window to start the flow for SOAP [11] that asked the user to select an IdP and what accounts the chat partner should prove access to (see Figure 1c). The user could choose as many of the given IdPs as they wanted and optionally fill in their communication partner's expected account names on these platforms. At the time of designing the study, the original prototype only supported Microsoft and Gitlab. To provide more providers, we omitted the technical procedure, and the journalists just responded with a predefined formatted string interpreted by Signal as a valid response on the participants' side. After receiving the response (Figure 1d), users could mark the user as verified, or, if an identifier was pre-filled, the client automatically marked the response as correct/incorrect.

The interface's UI can also be seen in Figures 4 and 5 in the Appendix. The source code is available at https://osf.io/dsyfr/.

Pre-Registration Mode The pre-registration mode is a variant of SOAP not proposed by Linker et al. but invented by us based on observations in the pilot study (see Section 3.2.6). Instead of asking the chat contact to prove access to a selected IdP and waiting for the response, the chat contact provided this proof in advance by logging in to the IdP once. This way, a new chat with this contact shows a SOAP response without any previous message exchange (see Figure 1a). This mode is similar to the proposal of Vaziripour et al. [24], where the participants liked that they did not have to communicate with the chat partner to check their identity. Technically, this would be possible in the same way the provider's server shares the public keys, or the material could be posted publicly as proposed by Vaziripour et al. [24].

3.2 Methodology

We conducted a lab study where we tested the detection rate of and the reactions to an impersonation attack on a new AC. The documents for the study can be found in the Appendices.

3.2.1 Setting and Scenario

When developing our scenario, we looked at previous studies on ACs. Herzberg at al. [5] reported that participants recognized to act differently depending on the situation, e.g., based

on the importance of a contact, and Wu et al. [27] discussed participants' need to be able to assess the need for an AC. Previous studies observing human behavior and ACs used very simple scenarios [18] or settings where there was little explicit (intrinsic or extrinsic) motivation for the participants to behave securely [5, 17, 24–27]. We wanted our participants to be motivated to conduct the AC, so we provided a scenario that gave them a reason to do so: a whistleblower scenario. We hoped the participants would understand the importance of being cautious, as they know the consequences of deanonymization, e.g., losing their job and reputation, prison, or even death. To check the realism of our scenario, we searched news sites and found examples where Signal was proposed as a channel for communication [1, 10, 15, 16, 23, 29].

In some previous studies with ACs, participants were invited in pairs [17, 26], sometimes knowing each other [26]; hence, they would have been able to judge whether the contacted person was the correct individual based on voice, looks, and behavior or meeting in person. We reduced these mitigating strategies through the scenario so the participants could not know the person they interacted with and could not verify the person via human characteristics.

Taking all this into consideration, we ended up with the following scenario outline: The participant, named Alex, is a whistleblower. Their colleague Hannah sent them documents revealing a political scandal via Signal. Their conversation was verified in person before receiving a .zip file containing sensitive data. Hannah is only reachable via Signal. Alex's task is to contact three journalists and send them the documents after ensuring they are interested in the data and the communication is safe. Alex receives information about these journalists on business cards (see Appendix A.6 for details). As part of the introduction, the participants were told that the business cards came from a trusted source. Communication could only occur through Signal's text function; other channels were not allowed. Each journalist had one intended possibility to be verified, for which we printed the necessary information on each business card: **Amira via safety number**, **Michael via QR code**, and **Anne via SOAP**. This way, the participants were nudged to use every method at least once. However, the participants were unaware that the authoritarian government of the scenario was suspicious of Alex and all connection attempts were attacked with impersonation attacks. Technically, this could be implemented by hacking the Signal servers and mounting a PITM, but also by possessing the SIM card, e.g., by a SIM swapping attack. So, all the verification checks failed: the safety number shown in Signal differed from the number and the QR code on the business card, and for each SOAP request, the provider or identity did not match (see Table 1). The only correct behavior for participants was to abort all communication attempts, which was explicitly allowed in the task description. The within-subject design allowed us to compare the available ACs and generate more observations with the small sample. We believed that

participants might buy into the scenario, but we want to note that although we described and explored a high-risk situation, the concrete setting lacked realism. We simplified by defining that the business cards Alex has available are to be trusted without exploring how realistic that is. Also, in a real-world scenario, we assume that Alex would compare the available information with further researched ones, e.g., an email address or a well-known social media account that can be confirmed from different sites. Also, deciding not to use Signal but to work via other channels is possible. To implement all possibilities realistically is sadly out of scope for a lab study and needs further research. More studies are needed to establish a best practice for real at-risk users. For this study, the most important was that the participants accepted the scenario as realistic and plausible.

SOAP Attackers Capabilities   The responses to the SOAP requests were randomly selected from three cases:
- Wrong provider - correct identifier
- Correct provider - wrong identifier
- Currently no access available

If multiple IdPs were asked for in a single request, the cases were picked without duplicates, so that with three asked IdPs, there were three different cases. The concrete available answers can be seen on Table 1.

Provider	Correct Identifier	Identifier available to the attacker
Amnesty.org	anne-baler-98746524b	anne-baler-13885412b
Facebook.com	Anne_Baler	AnneBaler
Gmail.com	n.a.	anne-baler_86@gmail.com
Twitter.com	@AnneBaler	@AneBaler
Amnesty.org	n.a.	a.patel@amnesty.com
Facebook.com	n.a.	Amira_Patel_86
Gmail.com	n.a.	patel_amira_86@gmail.com
Twitter.com	n.a.	@apatel
Amnesty.org	n.a.	m.kobel@amnesty.org
Facebook.com	n.a.	Michael_Kobel
Gmail.com	n.a.	michael_kobel@gmail.com
Twitter.com	n.a.	@michael_kobel

Table 1: This table displays the identifier the attacker sent and what the correct one would have been. Providers without correct identifiers are marked as "n.a.". For these, the participant could not determine the correct identifier. Participants received one of three responses: a) no identifier, simulating no current access to the account, b) an incorrect identifier for the requested provider, or c) a known identifier that is correct but for a provider different from the one requested.

SOAP does not submit the identifiers the requester expects. So, the attacker does not know whether the requester filled in identifiers. In the interface of the pilot study, if a provider was requested and the response did not contain the provider, it was marked as a missing provider. The attacker could send an accompanying message like "Sorry, I currently have no access to this account," hoping the requester would not mind. In the more opinionated later interface, any deviation from the request was marked as a failure. So, for the second half of the participants, we changed the attacker. The attacker would always send some form of identifier, hoping that the requester

had not filled in an expected identifier. If no identifier was filled in during request, the participants had to decide whether the identity submitted was sufficient.

3.2.2   Recruitment and Participants

We conducted a small pilot study (n = 4), recruiting participants from our research group's contacts. After this, we recruited 13 participants from an undergraduate usable security and privacy lecture and confronted them with an early interface. For their participation, the participants received bonus points for the lecture exam and a bonus cash reward. They started with €5, and if they securely transmitted the sensitive data to a journalist, they received an additional €5 for each journalist. The participants were told they lose everything if they got caught, e.g., by sending the data to the wrong person. As all journalists suffered an impersonation attack, no bonus could be earned. To eliminate any motivation to collaborate with fellow students, we paid each participant €20 and asked them to keep the study details confidential.

For the lab study, we recruited participants via a behavioral economics lab mailing list where studies can be distributed. To recruit 21 participants, an invitation was sent out to 3000[2] randomly picked mailing list receivers over 18 years old. The lab had a strict no-deception rule, so we had to change our initial reimbursement scheme. To keep a risk/reward payment scheme for motivation, the participants received a base pay of €15 and had the chance to receive an additional €9 (€3 for the correct decision for each journalist). The entire bonus cash reward would be lost if they made one wrong decision. We provided this reward to motivate the participants to contact as many journalists as possible and try the different authentication methods while behaving securely: weighing the risk of not sending the data and receiving less money versus sending the data and risking losing everything except the base pay. We hoped this would lead them to act cautiously and align their interests with the scenario. We followed that scheme with one exception: P2 did not send any message, being cautious that even a single message could be a problem, and thus stayed safe. To gather more information, we asked them to do so. While they later made a mistake, we paid out the bonus in full since their first behavior was safe.

In the lab study, eleven out of 18 participants did not use Signal before. Also, most of the participants (15) never checked the safety numbers of their contacts in any app. The ages ranged from 21 to 46 with a median of 24. One participant did not give their age. The pilot took place in July 2023, and the lab study in October 2023.

3.2.3   Ethics

We received IRB clearance for all studies and adhered to the German data protection laws and the GDPR in the EU.

---

[2]We had no control over how many people were contacted.

All participants consented to their participation and the use of the data for research purposes before participating. The participants were informed that they could terminate their participation at any time without negative consequences and that, in such a case, all the respective data collected up to that point would be deleted. The participants of the pilot study received bonus points for the lecture exam, which could also be obtained in other ways.

3.2.4 Study Protocol

The study was conducted in three parts, as described in the following.

Part 1 - Intro   The participants read and signed the consent form. Afterward, they received the material and were instructed to read the scenario text. Each participant was handed a pen, paper, and a smartphone with Android 13 and our modified version of Signal installed. Additionally, we handed them the three journalists' business cards (see Appendix A.6) in random order to counter ordering effects. The journalists each had an existing phone number to enable Signal communication. Further details on the business cards were fictive to avoid selection bias based on a newspaper's familiarity or reputation. The participants had to answer a quiz questionnaire on the phone before starting the scenario (see Appendix A.3). The quiz consisted of seven questions about the scenario. The participants could answer the questions as often as necessary to get all the answers correct.

Part 2 - Scenario   We asked the participants to think aloud while working on the task, audio recorded the whole procedure, and screen recorded the smartphone. Their task was to choose journalists and try to contact them securely. The researcher giving the briefing was present in the room during these steps and ended the scenario after about 30 minutes to keep the whole study under one hour. The researcher had the option to extend the time a few more minutes if a participant was in the final stage of sending or verifying. A second researcher who was not present in the room manned the journalists' Signal accounts. They had a playbook (see Appendix A.7) that was expanded in new situations. If a participant asked for a communication method other than Signal, this was denied, as is the case in real-world scenarios.

Part 3 - Outro   When a participant told the researcher they were done or the study time was up, they needed to complete a survey (see Appendix A.2). After this, there was a short interview followed by a debriefing (see Appendix A.5).

3.2.5 Analysis

We used qualitative and quantitative data to capture the results. As per our research questions, we were interested in:

1. Who tries to authenticate via a ceremony? We assumed this would be everyone as we added the red banner [25].
2. Which provider is chosen on SOAP? We assumed that most of the identifiers on the cards would be used.
3. Do participants detect the attack via SOAP? We assumed that most of them would.
4. How do the participants react? We assumed that the participants who detected a failed ceremony would abort contact.
5. How many participants fail the task? From the overall tone in related work, we assumed a few would.

A researcher who was present at all but one participant's sessions used notes, transcripts, screen recordings, and survey results to extract the steps participants took and where the participants failed. The researcher started by marking all positions in the recording relevant to the research questions, e.g., when a method was used and when and how a decision was made. A scenario is understood as failed if a participant sent a file to at least one journalist.

3.2.6 Results - Pilot study

In this section, we briefly describe the results of our pilot study. From the 13 participants (computer science students, abbreviated CS in the following), we excluded the data of three due to UI bugs and another participant who stated that they knew the study design beforehand. The data from the resulting nine participants was analyzed further. A table summarizing the results can be seen in the Appendix (Table 4).

The UI and the scenario text seemed to work, as all participants except CS-7 started every AC at least once.

Even though we intended for each journalist to be authenticated with exactly one method (safety number, QR code, or SOAP), all the business cards were provided with at least an email address. Following this and as we allowed to use custom providers, SOAP was not only used for Anne but for other journalists as well, with the work email being the most frequently used IdP (see Table 3 in the Appendix).

Overall, four of the nine participants forwarded the data to at least one journalist. All but one failure in the scenario can be traced back to SOAP. Specifically, we identified three reasons for failure.

Typosquatting: Three participants did not notice the typosquatting attack in SOAP or assumed it was acceptable, e.g., CS-3 recognized a provider mismatch but decided that an email identifier can only be verified as an account name if access to it is available. They all correctly saw that the safety number and QR code were invalid. We assume a more sophisticated attacker could have fooled more users.

"Marking" makes it secure: CS-7 contacted every journalist with a cover story. Afterward, they clicked on the safety number site, marked the journalists as verified, and sent the data. While that initially seemed rather strange to us, CS-7 explained in the interview and survey that they expected the chat to be verified and encrypted after this action.

Trust Chaining: Additionally, two of the participants verified one journalist and asked this journalist for the safety number of another journalist. The attacker provided the number seen by participants shown by the client. With this, the participants even accepted journalists who were previously perceived as suspicious.

Changes Based on the Pilot Study Based on the results, we made several modifications to our study design and how users interacted with the SOAP interface:

1. We adapted the SOAP interface to reduce possible attack surfaces and automated what could be automated.
2. We added a link to start a SOAP request on the safety number site in the settings.
3. The red banner no longer disappeared after verifying the person but turned green. This allowed a more direct way to the setting page and clearly indicated the chat's status.
4. To reduce the attack surface for typo attacks and match the current technical landscape, we removed the option to ask for custom IdPs and reduced the number of providers.
5. Based on the participants' comments and Vaziripour et al. [24], we assumed that a non-social media company would be favored and seen as more trustworthy. Therefore, we added Amnesty.org as a provider option.
6. We reworked the visuals of the UI, fixed glitches, added more text, and added guidance to the interaction of the SOAP responses depending on the outcome, e.g., obstacles to send in the case of an incorrect response.
7. As some of the participants in the pilot study were afraid of sending even a single message, they did not trigger the AC. To test SOAP without user interaction, we added the pre-registration mode as a between-subject condition (see Section 3.1.3), to which half the participants were assigned (see Table 2). We halved this group again by the provider/identity pair they would see: half saw the identifier for Facebook on Amnesty (Anne_Baler on Amnesty.org, condition "pre1"), and half saw a typo in the identifier (@AnneBaller on Twitter.com, condition "pre2"). We decided on this to get as many different perceptions as possible. We assumed the participants would most likely recognize the typo but might make a slip with the line on the business card and accept the incorrect assignment of identifiers.
8. To prepare for a more general, less tech-savvy sample, we rephrased "social authentication" as "proof of identity".
9. We changed the phrasing of the scenario, e.g., the reader was addressed more formally.
10. We made several smaller changes to the study documents and added the quiz section to ensure participants were at least once confronted with edge cases of the scenario.
11. When the scenario time was over, we asked the participants whether they wanted to make any further decisions.

3.3 Results

This section describes the lab study where we wanted to test the changes we made to the UI to prevent mistakes seen in the pilot study.

In general, seven out of 18 participants failed the task by sending the data to at least one journalist. Table 2 shows an overview of all the participants and to whom they sent the data. Most tried all available methods. The UI improvements generally prevented the mistakes observed in the pilot study. Nonetheless, the failure rate was still high. Below, we describe the results in detail.

3.3.1 Reasons for Failure

The scenario is considered a failure if a participant sends files to the impersonator. Seven participants failed the scenario for the following reasons: a) in-band safety number comparison (P8, P9, P10, P6), b) clicking too fast (P3), c) gambling for money (P13), and d) emotional stress (P2). The following paragraphs provide more details on those themes.

In-Band Safety Number Comparison The most common pitfall for participants was anchoring their trust in publicly known, unverifiable information, involving in-band exchanges of safety numbers.

P8 saw mismatches in the AC and asked Amira for her postal address and parts of the safety number. They decided that this was secret enough and sent the data. However, in the SOAP case, P8 stayed safe and decided against Anne because of an incorrect SOAP response.

P9 also saw the mismatches (QR, SOAP, safety number) but did not decide to stop and tried to find a way to communicate securely. They asked Michael why the scan failed. Michael said he reinstalled Signal and suggested sending the current chat's safety number, which was the attacker's and not the one on the business card. P9 agreed, compared, and marked the conversation as verified. After that, they tried to determine whether Amira was actually Amira by asking whether Amira knew them, as they assumed they had met when they exchanged business cards. Amira claimed to remember Alex and sent the chat's safety number within that communication. P9 also asked for the work address on the business card, and Amira reported the correct one. After that, Amira was marked as verified.

P9 saw the SOAP mismatch and asked Anne for a different way to verify her. Anna sent the current chat's safety number, but P9 was not entirely convinced, even though they marked Anna as verified. They noticed that some SOAP requests for social media profiles were still unanswered. At that point, P9 ran out of time and told the researcher their next step would be to send the material to Amira and Michael.

P10 was rather insecure and initially seemed overwhelmed by the scenario. They initially wanted to look at the data they

ID	Study Cond.	Sent to			Anne's IdPs				Study length	ATI [4]	Method attempted w/ journalists			Reason to Fail
		Anne	Amira	Michael	A	F	G	T			SOAP	QR	Safety	
P1	pre1	○	○	○	○	○	○	○	42 mins	3.9	-	M	Am	-
P2*†	ctrl	●	●	●	○	○	○	○	51 mins	3.4	-	M	-	Stress
P3†	pre2	●	○	○	○	○	○	○	36 mins	2.3	Am,M	-	-	Fast Clicking
P4	ctrl	○	○	○	●	●	○	●	44 mins	5.1	A	M	Am	-
P5	pre1	○	○	○	●	●	○	●	38 mins	4.3	A, Am	M	Am	-
P6†	ctrl	○	●	●	●	●	○	●	60 mins	3.4	A	M	Am,M	In-Band Comparison
P7	pre2	○	○	○	●	●	●	●	57 mins	4.0	A,Am,M	M	Am	-
P8†	ctrl	○	●	○	●	●	●	●	51 mins	3.0	A,Am,M	M	Am	In-Band Comparison
P9†	pre1	○	●	●	●	○	●	●	45 mins	5.6	A	M	A,Am,M	In-Band Comparison
P10†	ctrl	○	●	●	○	○	○	○	67 mins	3.2	-	M	Am,M	In-Band Comparison
P11	pre2	○	○	○	●	●	○	●	45 mins	3.1	A	M	Am	-
P12	ctrl	○	○	○	●	●	●	●	34 mins	3.8	A,Am	M	Am	-
P13†	pre1	○	●	○	●	○	○	●	52 mins	3.4	A	M	Am	Gambling
P14	ctrl	○	○	○	●	●	○	●	41 mins	5.2	A	M	A,Am,M	-
P15	pre2	○	○	○	●	●	○	○	42 mins	2.3	A	M	Am	-
P16	ctrl	○	○	○	●	●	○	●	36 mins	4.3	A,Am,M	M	Am	-
P17	pre1	○	○	○	○	●	●	●	47 mins	3.1	A	M	Am	-
P18	ctrl	○	○	○	○	●	○	●	32 mins	2.3	A,Am	M	Am	-

Table 2: Overview of the lab study participants' scenario results. Each "●" represents that the participant did what is depicted in the column, e.g., sent the data. The column "Anne's IdPs" marks which IdPs were requested by the participants. The names of providers and journalists are abbreviated (**A**mnesty, **F**acebook, **G**mail, **T**witter, **A**nne, **Am**ira, **M**ichael). The ∗ marks the participant who only continued the scenario after the researcher intervened. † marks participants who failed the scenario. The horizontal line after P11 marks the point where the attacker got stronger (see Section 3.2.1). More details such as the reasons for failure are discussed in Section 3.3.1.

received from Hannah, but as they had never used Android before, they got lost in the data management and needed help from the researcher to go back to Signal. They were told again that they did not need to look at the data for the scenario. They did not know what they should compare for Amira but managed to scan the QR code for Michael and recognize that this failed. Still, they told Michael they had sensible data and asked whether he could verify himself. Michael answered with the chat's safety number. At first, P10 was not sure how to compare the numbers but, after a while, realized that the sent number matched the security number in the settings. Afterward, the same happens with Amira. Anna was also asked for verification, but the scenario time was up.

P6 saw the mismatch for the QR code, safety number, and SOAP after requesting them. They asked Amira and Michael why the numbers were not correct. Both sent the current chat's number, and both received the data afterward. P6 told Anne that Signal said it was not secure to communicate based on the failed SOAP text. They even sent a screenshot when Anne said that this was not the case for her but did not send the data.

P3: Clicking Before Reading P3 was in the pre-registered SOAP condition. They saw a SOAP response without a request when they started a new chat with Anne. They clicked on "Mark Anne as verified" and sent her the data without recognizing this action because the chat was marked green and shown as verified. This makes P3 the only participant whose failure of the scenario is directly attributable to SOAP. After the interview, in which they stated that Anne had already been verified, they were presented with the video and were

surprised that they had actually clicked a button. They sent a SOAP request to Amira and Michael after seeing that they had to send a message for the other methods to work. They saw the faulty responses and deleted the chats afterward.

P13: Gambling for Money P13 was not sure who to send or not send the data to. After the time was up, they gambled and sent the data to Amira in the hope of getting more money. Although this is clearly related to the study design, we think it highlights that it was not clear to the participants what the secure and correct way to behave in this scenario was. On a similar note, another participant mentioned during the scenario and the interview that they thought it was strange that all the journalists were unsafe to send data to. They compared it to an exam situation where it seemed strange that all the questions had the same answer. Nevertheless, this participant behaved correctly. It appeared to require some effort for some participants to break off communication.

P2: Emotional Stress P2 initially decided not to contact any journalists, fearing that even sending a message would be too much. After the researcher intervened to tell them it would be all right, P2 went further. They saw the QR code mismatch and decided against SOAP requests, as they assumed they had to send an email and ended up confused. The participant read through the FAQs for safety numbers and ultimately decided to mark every journalist as verified, although they expressed being unsure of whether that was correct. Afterward, P2 sent the data. The participant was clearly highly emotional and insecure at that stage. In the interview, the participant expressed

frustration with their decision but stated they were emotional in the situation and could not think clearly. They were not aware of the safety number printed on the business card of Amira.

3.3.2 Study Conditions: Pre-Registered SOAP

We identified only one case where the condition negatively affected the results. Pre-registered SOAP failed once as P3 auto-clicked the decision. We think an obstacle, e.g., a time restriction or a different visualization, could have prevented that. Only two participants decided to send Anne the data, and they covered both conditions. Seven out of nine participants (pre-reg) sent another SOAP request. Although five of the seven clearly indicated being unsure or seeing the discrepancy. Only one participant did not communicate further with Anne.

3.3.3 Quantitative Data - Perception

It is difficult to compare the results with those from Vaziripour et al. [24] due to different scenarios, methods, and UIs. Nonetheless, with only their and our studies about SA available, we think it is sensible to point out similarities and differences between them. The scenario tested in our study did not involve any direct personal human contact other than through the chat. This was different in the study by Vaziripour et al. [24]. For example, participants could call each other for verification and meet in person to scan the QR code. The researchers found that their proposal of SA ranked higher than the other available methods in the Single-ease-question (SEQ). Conversely, we observed that the tested implementation of SOAP ranked lower than the other two methods provided (see Figure 2b in the Appendix). Potentially in relation to the other available methods, SA ranked much lower in Vaziripour et al.'s study in the trust score than the other extremely high-ranking methods. We observed a mix of perceptions (see Figure 2c). The participants in our study generally trusted all the methods less than the participants in Vazirpour et al.'s study [24]. However, SA still ranked the lowest in both studies. We also asked participants whether they were confident in their decision with the method and saw that SA ranked third in this category while only leading to one failure in the scenario.

3.3.4 Participants' Perceptions and Understandings

We briefly interviewed each participant, asking them about their understanding of the ACs. We found that only two participants had a detailed understanding of safety numbers and the QR code. They used terms like "E2EE" and "public/private keys". One of them studied computer science, where they learned about this, and the other person recognized parallels from email encryption. Four other participants mentioned terms like "E2EE" but had no further concepts of it. They have heard the terms before and connected them to Signal.

Although their technical knowledge was limited, many participants conducted the ACs correctly. All participants understood that the QR code, safety number, or accounts should have matched with what was given. All four participants who asked for the safety number via chat detected a mismatch in the QR code or safety numbers beforehand. Only P10 did not see the safety number on the business card. The four participants asked for the number in the chat as a mitigation tactic. All of them were convinced that it is safe to send after receiving the current chat's safety number (see "in-band comparison" in Section 3.3.1).

SOAP, or "proof of identity" as it was called in the study, was known to no one. Speculations on how SOAP worked were, similar to the other ACs, very vague. Often, the participants only stated how they used it and that the accounts should have matched. The participants believed there must be some kind of connection between the accounts provided and the Signal account. It was speculated that this could be based on a one-time token that must be entered (similar to SMS codes that are sent if a phone number is used as an account name), that a person needs to add the number to the account at the IdP, or more generally that the journalists need to log in into the account and do something. Another belief we encountered was that SOAP was based on a setup that happened during the Signal account creation. No participant mentioned safety numbers in their ideas about SOAP.

The participants thought that if the journalists had to take action to create a response, a typo in the account name could occur. Therefore, the participants requested SOAP multiple times to rule out such cases, just as they scanned the QR code multiple times.

According to Section 3.3.3, the participants trusted SOAP less than the other methods. In the interviews, one participant was confused that SOAP gave a valid response, although the QR code was mismatched. Also, the participants thought that account names could somehow be faked or an IdP could be hacked. We additionally found that account names were perceived as private and that the participants were unsure about the processes occurring on the side of the to-be-verified person.

4 Discussion

We conducted a lab study with 18 participants to observe social authentication (SA), an authentication ceremony (AC), in a no-win attack scenario. In this section, we discuss our results from the perspective of our research questions.

4.1 RQ1 & RQ2: Resistance Against Impersonation Attacks - Detection and Reaction

This study observed a SA ceremony in an attack scenario. We were interested in how resistant SA would be against impersonation attacks. So, as a first step, we researched an

attacker who used typo squatting attacks to impersonate the communication partners. While with the simple interface in the pilot study, three participants failed because of SOAP, only one participant in the pre-registered condition failed in the lab study. The UI heavily supported the participants in detecting mismatches when an identifier was given. We applied a strong, opinionated design, e.g., interpreting anything other than the requested identities as incorrect and reducing possible providers to a fixed list. That seemed to help, but we could not measure long-term effects in our setting.

The participants often tried one method, then tried another, changed the journalist, returned to the first, and sometimes retried a previous method. The participants' flow through the tasks was not linear. Not all the participants reacted as hoped to an incorrect SOAP response. Some of the **participants retried** SOAP after seeing an incorrect response or even tried further and asked via text for a way to authenticate the other person. While we found plausible reasons for the first case, we cannot rule out that it is a study artifact. We think this should be investigated further in future studies and considered when designing studies and interfaces. The participants not aborting the communication does not necessarily reflect the hope connected to SA: an intuitive method for recognizing whether you are communicating with the right person. It is also likely that the lab setting influenced the participants, e.g., through demand effects. We, therefore, understand the results as an upper bound for failures in ACs.

The participants without any technical knowledge about what happened **concluded that something was wrong**, although they did not necessarily attribute this to a malicious actor, despite the explicit mention of them in the scenario. Inputting the identifier beforehand helped the automatic detection and, therefore, the automated decision. Based on our data, we do not know whether this can be expected in a real scenario. It is, e.g., unclear where users would source the identifiers from. Anecdotally, in the pilot study, a participant was unsure whether there were unique Facebook identifiers and where to find them. There are paths to help the user here, e.g., if the person's identifier is not known, external means can verify it afterward (e.g., seeing connections in a social graph or validation through a third site). We think there are many possibilities for how this can develop over time, and it is an important area for future work.

While we think what we observed is promising, the sample was too small to draw strong conclusions regarding resistance against impersonation attacks in the real world.

Safety Numbers, on the other hand, did not seem to be resistant to impersonation attacks. While safety numbers were not the focus of the study, we want to highlight that some of the participants failed the scenario because they did an in-band exchange and comparison of key material. As safety numbers comparison is a currently available AC, this should be researched further, as well as whether this has a negative real-world impact. We suggest seeing whether some preven-

tive action can be taken on the client side, e.g., by pattern matching and informing users when they attempt to exchange safety numbers via chat.

4.2 RQ3: Perception and the Role of the Identity Provider

Although only one participant made a mistake with SOAP, the participants were not as confident about their decisions with SOAP as with the other methods. The same trend existed for the perceived usability or trustworthiness of the method to verify their contact. However, the small failure rates contradict that perception. We argue this could be a positive situation for SA. The usability aspect seems solvable, and the participants behaved as intended. But, for the other methods, they behaved insecurely but felt as confident as with SA, creating an "illusion of security" [5]. Regarding SOAP, the participants behaved as hoped. Now, we need to improve the participants' confidence in their own judgment based on SOAP. We are not sure where the difference in the perception of the methods comes from. The sample was too small to make any sensible statistical inferences, but we think further research could investigate the phenomenon.

4.3 Further Observations

This section covers themes beyond our research questions that may offer relevant insights to researchers and practitioners.

Identification of the Person vs. Authentication of the Connection Similar to other studies [3], we observed that the participants did not fully understand how encryption works and, following this, what an attack would look like. We observed, e.g., the assumption that if you have the correct phone number, you will end up with the correct person. In combination with the theme of the "almighty hacker"(see Dechand et al. [3]), participants assumed there is nothing a user can do to protect their communication effectively. So, explaining to the participant that doing something is necessary to communicate with the correct person may be easier than explaining that something is necessary to prevent others from listening. In short, the mental models of Signal's functions did not seem to align enough with the technical reality to understand an attacker. Considering this, it is understandable why the participants fell back to using addresses and shared secrets, or something perceived as such, to identify the other person.

4.4 Protocol/UI Challenges

Multiple requests are not a problem for the protocol per se but can complicate the UI. When designing a protocol and the corresponding interface, designers should remember that the interaction may involve multiple, sometimes canceled, requests. For example, on the one hand, we wanted to ensure that no data were sent with a failed request, but on the other

hand, a typo in the expected identifier was possible and needed to be traceable (false-negative). The participants wanted to believe the other person was legitimate. They were looking for a way to send the data rather than a reason not to send it.

To get the safety number or to receive a SOAP response, a chat contact has to **communicate with the other party**. Depending on the scenario, this communication can be problematic, and participants may hesitate to communicate. If the server is trustworthy, one can reduce the friction here. However, if one also does not trust the server, this is still a problem to be solved. For future studies, communication can be explicitly allowed in the scenario to reduce participants' confusion. In the study, pre-registration caused one failure but helped participants identify issues in other cases.

Vaziripour et al. [24] concluded that the necessary infrastructure for SA "needs to be more trusted than social media companies". We observed that the participants wanted to ask for the journalists' working email addresses. Such **custom providers** are not intended by the (SOAP) protocol. Allowing custom providers also allowed typo attacks on the provider level, making everything even more complicated. It is necessary to determine whether the usage of SA can be reduced to a fixed set of providers, depending on the use case. For example, in a company, setting the list of providers could vary vastly from that of instant messaging for personal use. We suggest finding a way to allow additional, possibly ad-hoc selected providers without impacting security. With SOAP, Linker et al. [11] proposed an interactive communicative way to verify a person. Vaziripour et al. [24] proposed an asynchronous interaction with the previously made public public key. We simulated this in the pre-registered condition after we observed that participants hesitated to even write a single message before verifying a person. We think this hesitation will not appear in most scenarios, but for those where it matters, pre-registration solves a problem. We thus suggest investigating further how an asynchronous solution could be achieved or how the interactive solution can reduce friction.

4.5 Signal Specifics

Some observations made are highly specific to the Signal app and may spark discussions about the UI. Some participants were confused by the way the safety numbers were presented. If a user opens the safety number page, the numbers appear in an animation, giving the impression they are generated just then and would change every time the site is visited.

Signal has the option to use the camera from the start screen. The participants tried using this feature to scan the QR code of the safety number. Here, direct feedback that something was done incorrectly or what type of data might have been scanned could have helped the participants. Signal allows safety numbers to be compared from the clipboard, but no participant was aware of that. When something that looks like a safety number appears within a chat, Signal can provide

additional information, e.g., to prevent in-band comparison, but also enhance the sharing of fingerprints between already verified contacts. Similarly to Shirvanian et al. [18], we observed situations where the participants had to compare long numbers across multiple views, but that was not intended and insecure to do in our scenario.

5 Limitations

Conducting a lab study comes with limitations. Participants may behave differently than they would in real life. For our study, this could have led to more interaction and attempts even if the participants thought they should stop. The used reward and risk system is not the same as being a whistleblower and getting caught by the government. But unlike previous studies, which had no risk, we offered a real tradeoff. However, it is still a role play, and we are unaware of the extent of the impact. The setting of a lab study might lead participants to continue because they think there must be a way. We ensured that participants knew that all the connections were potentially insecure and that no communication was also an option. Also, within the scenario, trying different methods to authenticate the journalists was unproblematic. Nonetheless, feedback suggested that the participants liked the scenario and tried to empathize with the situation. We had to pick a fixed set of providers. To not only rely on U.S.A.-based providers, we added Amnesty.org, although it does not offer an identity service to work with SOAP. We do not believe any participant knew this technical detail. Due to a bug with Amira, some of the participants did not need to send a message to get the safety number. We saw that sending a message made the participants hesitant, but ultimately, they all decided this was not a show-stopper.

6 Conclusion

To test a new social authentication (SA) protocol called SOAP and compare it with traditional ACs, we developed a scenario-based lab study where participants take over the role of whistleblowers and try to gauge whether the connections to journalists contacted via the Signal app are secure. Based on a pilot study, we improved an interface for SOAP and made it similarly usable as manual safety number comparison or QR codes. We found that although the participants did not know how SA worked, they behaved mostly securely, and mistakes were more often made in existing ACs. These findings make us optimistic about SA as a usable AC. While our sample size was rather small, and our scenario may not directly translate to a realistic real-world situation (e.g., at the current time, we do not recommend whistleblowers to use SOAP), it provided the participants with understandable reasoning and motivated them to act securely. With the study design, we provide a template for further research and comparison of ACs.

7 Acknowledgements

We are grateful to our shepherd and the anonymous reviewers for their valuable comments and suggestions. We thank the Werner Siemens-Stiftung (WSS) for their generous support of this project.

References

[1] Whistleblower Aid. Become a whistle-blower. https://whistlebloweraid.org/become-a-whistleblower/signal/. Accessed: 30 October 2023.

[2] Mashari Alatawi and Nitesh Saxena. SoK: An Analysis of End-to-End Encryption and Authentication Ceremonies in Secure Messaging Systems. In *Proceedings of the 16th ACM Conference on Security and Privacy in Wireless and Mobile Networks*, WiSec '23, page 187–201, New York, NY, USA, 2023. Association for Computing Machinery.

[3] Sergej Dechand, Alena Naiakshina, Anastasia Danilova, and Matthew Smith. In Encryption We Don't Trust: The Effect of End-to-End Encryption to the Masses on User Perception. In *2019 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 401–415, June 2019.

[4] Thomas Franke, Christiane Attig, and Daniel Wessel. A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale. *International Journal of Human–Computer Interaction*, 35(6):456–467, April 2019.

[5] Amir Herzberg and Hemi Leibowitz. Can Johnny finally encrypt?: Evaluating E2E-encryption in popular IM applications. In *Proceedings of the 6th Workshop on Socio-Technical Aspects in Security and Trust*, pages 17–28, Los Angeles California, December 2016. ACM.

[6] Amir Herzberg, Hemi Leibowitz, Kent Seamons, Elham Vaziripour, Justin Wu, and Daniel Zappala. Secure Messaging Authentication Ceremonies Are Broken. *IEEE Security & Privacy*, 19(2):29–37, March 2021.

[7] Sakshi Jain, Neil Zhenqiang Gong, Sreya Basuroy, Juan Lang, Dawn Song, and Prateek Mittal. New Directions in Social Authentication. In *Proceedings 2015 Workshop on Usable Security*, San Diego, CA, 2015. Internet Society.

[8] Keybase. Keybase book. https://book.keybase.io/docs/server#meet-your-sigchain-and-everyone-elses. Accessed: 8 January 2024.

[9] Sean Lawlor Lewi, Kevin. Deploying key transparency at WhatsApp. https://engineering.fb.com/2023/04/13/security/whatsapp-key-transparency/, April 2023.

[10] Guardian News & Media Limited. How to contact the guardian securely. https://www.theguardian.com/help/ng-interactive/2017/mar/17/contact-the-guardian-securely. Accessed: 30 October 2023.

[11] Felix Linker and David Basin. Soap: A social authentication protocol. https://arxiv.org/abs/2402.03199, 2024.

[12] Lee Livsey, Helen Petrie, Siamak F. Shahandashti, and Aidan Fray. Performance and Usability of Visual and Verbal Verification of Word-Based Key Fingerprints. In Steven Furnell and Nathan Clarke, editors, *Human Aspects of Information Security and Assurance*, volume 613, pages 199–210. Springer International Publishing, Cham, 2021. Series Title: IFIP Advances in Information and Communication Technology.

[13] Marcela S. Melara, Aaron Blankstein, Joseph Bonneau, Edward W. Felten, and Michael J. Freedman. CONIKS: Bringing key transparency to end users. In *24th USENIX Security Symposium (USENIX Security 15)*, pages 383–398, Washington, D.C., August 2015. USENIX Association.

[14] moxie0. Safety number updates. https://signal.org/blog/safety-number-updates/. Accessed: 2 January 2024.

[15] ZEIT ONLINE. Appell an potenzielle whistleblower. https://www.zeit.de/administratives/2019-01/technologiebranche-whistleblower-suche. Accessed: 30 October 2023.

[16] The Washington Post. Submit an anonymous news tip. https://www.washingtonpost.com/anonymous-news-tips/. Accessed: 30 October 2023.

[17] Svenja Schröder, Markus Huber, David Wind, and Christoph Rottermanner. When SIGNAL hits the Fan: On the Usability and Security of State-of-the-Art Secure Mobile Messaging. In *Proceedings 1st European Workshop on Usable Security*, Darmstadt, Germany, 2016. Internet Society.

[18] Maliheh Shirvanian, Nitesh Saxena, and Jesvin James George. On the Pitfalls of End-to-End Encrypted Communications: A Study of Remote Key-Fingerprint Verification. In *Proceedings of the 33rd Annual Computer Security Applications Conference*, ACSAC '17, pages

499–511, New York, NY, USA, December 2017. Association for Computing Machinery.

[19] Signal. Signal messenger: Speak freely. `https://signal.org/`. Accessed: 13 February 2024.

[20] Joshua Tan, Lujo Bauer, Joseph Bonneau, Lorrie Faith Cranor, Jeremy Thomas, and Blase Ur. Can Unicorns Help Users Compare Crypto Key Fingerprints? In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3787–3798, Denver Colorado USA, May 2017. ACM.

[21] The Telegram Team. Colorful calls, thanos snap effect, and an epic update for bots. `https://telegram.org/blog/calls-and-bots`. Accessed: 2 January 2024.

[22] Telegram. Faq for the technically inclined. `https://core.telegram.org/techfaq#man-in-the-middle-attacks`. Accessed: 8 January 2024.

[23] The New York Times. Got a confidential news tip? `https://www.nytimes.com/tips`. Accessed: 30 October 2023.

[24] Elham Vaziripour, Devon Howard, Jake Tyler, Mark O'Neill, Justin Wu, Kent Seamons, and Daniel Zappala. I Don't Even Have to Bother Them! Using Social Media to Automate the Authentication Ceremony in Secure Messaging. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–12, New York, NY, USA, May 2019. Association for Computing Machinery.

[25] Elham Vaziripour, Justin Wu, Mark O'Neill, Daniel Metro, Josh Cockrell, Timothy Moffett, Jordan Whitehead, Nick Bonner, Kent Seamons, and Daniel Zappala. Action needed! helping users find and complete the authentication ceremony in signal. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, pages 47–62, Baltimore, MD, August 2018. USENIX Association.

[26] Elham Vaziripour, Justin Wu, Mark O'Neill, Jordan Whitehead, Scott Heidbrink, Kent Seamons, and Daniel Zappala. Is that you, alice? a usability study of the authentication ceremony of secure messaging applications. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*, pages 29–47, Santa Clara, CA, July 2017. USENIX Association.

[27] Justin Wu, Cyrus Gattrell, Devon Howard, Jake Tyler, Elham Vaziripour, Daniel Zappala, and Kent Seamons. "Something isn't secure, but I'm not sure how that translates into a problem": Promoting autonomy by designing for understanding in Signal. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, pages 137–153, 2019.

[28] Tarun Kumar Yadav, Devashish Gosain, Amir Herzberg, Daniel Zappala, and Kent Seamons. Automatic detection of fake key attacks in secure messaging. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, CCS '22, page 3019–3032, New York, NY, USA, 2022. Association for Computing Machinery.

[29] Süddeutsche Zeitung. So erreichen sie das investigativ-team der süddeutschen zeitung. `https://www.sueddeutsche.de/projekte/kontakt/#messenger`. Accessed: 30 October 2023.

A   Appendix

The original study material in German can be found online at `https://osf.io/dsyfr/`. Due to space constraints, we have only included the translated versions in this paper.

A.1   Scenario

This are the translated scenario texts available to the participants, including the payment description for both studies.

A.1.1   Scenario Text for the Pilot Study

The study consists of a role play in which you take on the role of Alex. The scenario is described in the following text. Please read the text carefully and put yourself in the situation.

Scenario card Your name is Alex and you live in a country ruled by an authoritarian regime. Both blanket and targeted surveillance is a daily phenomenon. You work in a high-ranking government agency. A colleague, Hannah, has gained access to extremely sensitive information about high-level corruption and shared it with you in encrypted form through the Signal app a few months ago. This information includes revelations about illegal activities by politicians.

You want this information to be made public. In order to avoid drawing suspicion to Hannah, who had access to the data, you have decided to wait a few months and then send the data to journalists. The time has now come and you can begin.

You already have business cards from three trustworthy investigative journalists from abroad. You have received them personally and you trust the information on them. All journalists are known for their integrity and have already uncovered a number of major scandals. All three journalists offer whistleblowers that they can be contacted securely via the Signal app.

You have a rough idea of how such a contact works: First you send the data to the journalist. The journalist then does research and checks whether the data is genuine. Once they are satisfied, they publish the story. This can take a while. The archive containing the data and explanation can be found in your Signal app in the chat with Hannah. You are familiar with the content and the exact content does not matter for the study.

Your goal is to ensure that all three journalists receive the data about the corruption. Considering the dangers you and Hannah face if your government's intelligence agencies find out that you have leaked the data, it is crucial for you to make sure that you communicate with the journalists in encrypted form using the Signal app. You are sure that as long as you use the Signal app correctly, the secret services will not be powerful enough to break the encryption or access the metadata.

Bonus payment: You currently have 5€ in your account. For every journalist you successfully send the data to, you will receive another €5.

However, if you are caught by the secret service, you will end up in prison and will not receive any payment. So only send the data if you are sure that the Signal app will protect you. You will receive the exam bonus of 2% points even if you end up in prison. If you are not caught, you will receive the 2% points and the money from your account.

Instructions

Signal Signal is an encrypted messenger and phone app. Signal saves your number, but does not create a log file for your incoming or outgoing communication. Signal is easy to use: Open the app and tap the pencil icon (bottom right on Android phones) to write a new message. Enter the desired phone number in the search field. You can now send an encrypted message via Signal.

How do I take screenshots? Press and hold the "On/Off" button and "Volume down" button on your phone at the same time for about one second.

A.1.2 Scenario Text for the Lab Study

The study consists of a role play in which you take on the role of Alex. The scenario is described in the following text. Please read the text carefully and put yourself in the situation.

Scenario Description

Your name is Alex and you live in a country ruled by an authoritarian regime. Both blanket and targeted surveillance happen on a daily basis. You work in a high-ranking government agency. A colleague, Hannah, has gained access to extremely sensitive information about high-level corruption and shared it with you in encrypted form through the Signal app a few months ago. This information includes revelations about illegal activities by politicians. You want this information to be made public. In order to avoid drawing suspicion to Hannah, who had access to the data, you have decided that you will wait a few months and then you (Alex) will send the data to journalists. The time has now come and you can begin.

You already have business cards from three trustworthy investigative journalists from abroad. You have received the business cards personally from the journalists and you trust the information on them. All journalists are known for their integrity and have already uncovered a number of major scandals. All three journalists offer whistleblowers that they can be contacted securely via the Signal app.

You have a rough idea of how such a contact works: First you send the data to the journalist. The journalist then researches and checks whether the data is genuine. If the journalist is convinced, the story is published. This can take a while. The archive containing the data and explanation can be found in your Signal app in the chat with Hannah. They are familiar with the content but the exact content does not matter for the study.

Your goal is for all three journalists to receive the data on corruption. Considering the dangers you and Hannah face if your government's intelligence services find out that you have leaked the data, it is crucial for you to ensure that you communicate with the journalists in encrypted form using the Signal app. You are sure that as long as you use the Signal app correctly, the intelligence services will not be able to break the encryption or access the metadata.

Payment: You will receive a basic payment of €15 after completing the study. You also have the option of receiving a bonus of up to €9.

You should send the data to the journalists with a secure connection - and only to those with a secure connection.

A decision must be made for each of the three journalists individually:
- If the connection is secure, the data must be sent.
- If the connection is insecure, no data may be sent.

For each correct decision you receive a €3 bonus, i.e. up to €9 in total. But: No bonus is awarded, - if data is sent via at least one insecure connection - or if no data is sent although there is at least one secure connection.

How to send messages with Signal

Open the app and tap the pencil icon at the bottom right to write a new message. Enter the desired phone number in the search field. You can now send an encrypted message via Signal.

What is a secure connection?

The Signal app offers you methods to ensure that you are communicating with the right person and correctly encrypted. If you cannot use the app to ensure that the connection is secure, you should assume that the connection is insecure.

A.2 Survey

The survey varied slightly in the pilot and the lab study. Social authentication was called "Proof of identity (Identitätsnachweis)" in the lab study and participants were addressed more formally. The questions that were exclusively part of a study or edited a lot are marked.

Q1: Below are some questions about the methods you interacted with during the study. The lab study includes a role play. However, please do not fill out this questionnaire in the role of Alex, but as yourself. (Type: Text)

Q2: Please enter your study pseudonym (Type: Text Entry)

Q3: Do you use the Signal app independently of the study? (Type: MC) Answer Choices: *"No", "Yes, Rarely", "Yes, Often"*

Q4: Before you took part in the study: For how many of your chat contacts did you use a safety number (e.g. in Whatsapp or Signal) to verify the contact? (Type: MC) Answer Choices: *"With none of my chat contacts", "With some of my chat contacts", "With about half of my chat contacts", "With most of my chat contacts", "With almost all of my chat contacts".*

Q5: During the study, you tested up to three different methods of verifying a contact via Signal. a) via QR code scan b) comparing safety numbers c) via account affiliation on platforms (social authentication). The following questions are about your thoughts on exactly these methods. (Type: Text)

Q6: Which of the methods did you use in the course of the study? (Type: MC) Answer Choices: *"QR code", "safety number", "social authentication"*

Q7: How much do you agree with the following statement: I have confidence in this method of verifying safety numbers in Signal. *Lab study: I have confidence in this method for verifying the identity of my conversation partners.* (Type: Matrix) Items: *"safety number", "social authentication", "QR code"*
Scale (5): *Strongly disagree, Somewhat disagree, Neither agree nor disagree/neutral, Somewhat agree, Strongly agree*

Q8: How much do you agree with the following statement: I am sure that I made the right decision when using the method. (Type: Matrix) Items: *"safety number", "social authentication", "QR code"*
Scale (5): *Strongly disagree, Somewhat disagree, Neither agree nor disagree/neutral, Somewhat agree, Strongly agree*

Q9: In terms of verifying with the appropriate method: Overall, how difficult or easy was it to complete the task? *Lab study: Related to verifying identity using the appropriate method: How did you find completing the task* (Type: Matrix) Items: *safety number, social authentication, QR code*
Scale (5): *Very difficult, Very easy*

only lab study: Q10: Please mark which method you would choose if you had to verify a friend. (Type: MC) Items: *"safety number", "social authentication", "QR code"*

Q11: Is there anything else you would like to tell us about the methods? (Type: Text Entry)

Q12: Please indicate your level of agreement with the following statements. (Type: Matrix) Items: *"I understood the scenario", "I think the scenario is plausible", "I thought myself into the scenario", "The chance of getting money motivated me to contact as many journalists as possible", "The risk of losing money motivated me to be careful", The financial incentive helped me to empathize with the scenario", "Without a financial incentive I would not have taken the scenario so seriously", "Without a financial incentive I would not have gone to so much trouble to check the security numbers".*
Scale (5): *Strongly disagree, Somewhat disagree, Neither agree nor disagree/neutral, Somewhat agree, Strongly agree*

Q13: The following is about your interaction with technical systems. By 'technical systems' we mean apps and other software applications as well as complete digital devices (e.g. cell phone, computer, TV, car navigation). Please indicate your level of agreement with the following statements.

(Type: Matrix) Items: *"I like to take a closer look at technical systems", "I like to try out the functions of new technical systems", "I primarily deal with technical systems because I have to", "When I have a new technical system in front of me, I try it out intensively", "I like to spend a lot of time getting to know a new technical system", "It is enough for me that a technical system works, I don't care how or why", "I try to understand exactly how a technical system works", "It is enough for me to know the basic functions of a technical system", "I try to make full use of the possibilities of a technical system".*

Scale (6): *Not true at all, Not true to a large extent, Rather not true, Rather true, Moderately true, Completely true*

Q14: How old are you? (Type: Text Entry) *only lab study:*

Q15: Which gender do you feel you belong to? (Type: MC) Answer Choices: *"Female", "Male", "Diverse", "I would like to describe myself:", "Not specified"*

Q16: Which employment situation suits you? What in this list applies to you? Please note that gainful employment is understood to mean any paid or income-related activity associated with an income. (Type: MC) Answer Choices: *"Full-time employment", "part-time employment", "partial retirement (regardless of whether in the working or release phase)", "marginally employed, 450-euro job, mini-job", ""one-euro job" (in receipt of unemployment benefit II)", "occasionally or irregularly employed", "In vocational training/apprenticeship", "In retraining", "Voluntary military service", "Federal voluntary service or voluntary social year", "Maternity leave, parental leave, parental leave or other leave of absence (click on the relevant option for partial retirement)", "Not gainfully employed (including: Pupils or students who do not work for money, unemployed, early retirees, pensioners without additional income)"*

Q17: If you are not in full-time or part-time employment: Please say, which group on this list you belong to. (Type: MC) Answer Choices: *"Pupils at a general school", "students", "pensioners, retired, early retired", "unemployed", "permanently disabled", "housewives/househusbands", "other, namely:"*

Q18: Please note that it is important that the questions asked in this questionnaire are answered by each participant independently and without prior knowledge of the study. This ensures the integrity and quality of our data. We therefore ask you not to share any information about the content of the study or the questions of this questionnaire with other people for 2 weeks and your answer to the next question will have no effect on you, your bonus points or bonus payment! But it is very important for us that you answer honestly. Did you already know the details of what happens in the study before participating in the study? *Lab study:*Please note that it is important that the questions asked in this questionnaire are answered by each participant independently and without prior knowledge of the study. This ensures the integrity and quality of our data. We therefore ask you not to share any information about the content of the study or the questions in this questionnaire with other people for one week. Your answer to the next question will not affect you or the money you receive at the end of the study! But it is very important to us that you answer honestly. Before participating in the study, did you already know details about what will happen in the study? (Type: MC) Answer Choices: *"Yes", "No"*

Q19: What did you know and how do you think it affected you? (Type: Text Entry)

Q20: Thank you for completing the questionnaire. Now please turn to the person in the room. (Type: Text)

A.3 Quiz

In the lab study the participants had a to complete a quiz after reading and before starting the scenario. They could answer questions as often until they had all correct.

Q21: The following questions are intended to ensure that you have carefully read and understood the assignment. You can use all available documents to answer the questions.

Q22: What is the name of the person you are supposed to play? (Type: MC) *Answer Choices: "Alex", "Hannah", "Friedrich", "Eva"*

Q23: What should you do if you have established a secure connection with a journalist? (Type: MC) *Answer Choices: "Send the data to the journalist and try to contact other journalists", "Cancel the contact", "Let Hannah know", "Send the data to the journalist. The task is then completed."*

Q24: What should you do if you cannot ensure that a connection to a journalist is secure? (Type: MC) *Answer Choices: "Cancel the contact", "Send the data to the journalist", "Let Hannah know".*

Q25: Under what conditions should you send the data to whom? (Type: MC) *Answer Choices: "All journalists, even if I can't be sure that the connections are secure", "Every journalist with whom there is a secure connection", "Hannah".*

Q26: In which situations does the bonus payment increase? (Mehrfachnennung ist möglich.) (Type: MC) *Answer Choices: "A connection is not secure and I am not sending data", "A connection is secure and I am sending data", "A connection is not secure and I am sending data", "A connection is secure and I am not sending data".*

Q27: What possible situations can occur in the study? (Mehrfachnennung ist möglich.) (Type: MC) *Answer Choices: "All connections are secure and I send the data to all journalists", "No connection is secure and I don't send the data to anyone", "Some connections are secure and I send the data there".*

Q28: Which statement is true? (Type: MC) *Answer Choices: "The information on the business cards is correct", "The information on the business cards may be incorrect".*

A.4 Interview Guideline

1. Why did you decide to act the way you did with the journalists? (Go through it step by step, was impersonation a conceivable option?)

2. How do you think the methods work? (safety number, QR code, social authentication)

3. Do you have an idea where you would like to apply such a method?

4. Which method would you use if you had to verify a friend? (Focus on why)

5. Would you be willing to use your accounts for social authentication?

6. Would you behave differently as a whistleblower outside of the study?

A.5 Debriefing Guideline

1. Have the payment form filled out.

2. Ask the participant not to talk about the study for one week. Explain how things work would render the data unusable.

3. Explain the objectives: To see if an impersonation attack is detected and what the reactions are. We were also interested in what are the thoughts concerning the procedure.

4. Explain: Security numbers must come through a different channel than the conversation. They change if there is an eavesdropper, but also if, for example, one changes their phone and reinstalls Signal.

5. Are there any questions?

A.6 Business cards

This is the information on the business cards participants had available. All the information except the phone numbers were made up.

Amira

- Amira Patel
- Investigative journalist
- Hallentorstraße 4, 20654 Hamburg
- Phone: {removed as the numbers actually exist}
- Mail: amira_patel@newsorg.de
- Signal Safety Number: 72500 10336 57813 26686 75084 04894

Anne

- Anne Baler
- Investigative journalist
- Isarwege 15, 80542
- Phone: {removed as the numbers actually exist}
- Mail: anne_baler@newsunion.de
- Twitter.com: @AnneBaler
- Facebook.com: Anne_Baler
- *(only pilot study)* Instagram.com: @AnneBaler
- *(only pilot study)* LinkedIn: anne-baler-98746524b
- *(only lab study)* Amnesty.org: anne-baler-98746524b

Michael

- Michael Kobel
- Investigative journalist
- Maurerweg 28, 10235 Berlin
- Phone: {removed as the numbers actually exist}
- Mail: michael_kobel@newsbody.de
- {Image of a QR Code}

A.7 Playbook

This is the playbook the researcher, acting as the journalists, used. It was extended when necessary, and the wording varied slightly from case to case.

Participant: Hello

Journalist: Hello|Hi|Good morning|Good afternoon|Good evening|Good night

Participant: Hello, my name is Alex

Journalist: Hi, I'm [Person Name]

Participant: Hi, do you still know me?

Journalist: No, I don't know whose number this is.

Participant: I got your business card some time ago.

Journalist: I give these to a lot of persons. Sorry that I cannot recall you.

Participant: It's me, Alex.

Journalist: Ah, I can briefly remember.

Participant: I cannot verify your safety number on the business card. Do you have another idea?

Journalist: Yes, my new number is [safety number].

Participant: Can you verify yourself?

Journalist: Yes, I can send you my current safety number.

Participant: Your safety number is wrong

Journalist: My number has changed. My new one is [safety number of conversation].

Participant: Your QR Code does not match.

Journalist: I lost my old phone and had to buy a new one.

Participant: Do you have a [safety number|QR Code] for me?

Journalist: [Current safety number|QR Code sent]

Participant: I cannot contact you because Signal says, there is a chance that the conversation is not secure

Journalist: I don't see this message.

Participant: Screenshot of failed SOAP message

Journalist: My accounts were hacked before and I had to create new accounts. I don't have new ones on [missing Providers here]. On [changed Username providers], I have a new user handle.

Participant: How does your [email, address, phone number, other PII] look like?

Journalist: [Publicly available data]

Participant: I would like to send you the data

Journalist: Okay, I'm ready

Participant: [data sent]

Journalist: Thank you! I will check the files with my colleagues and get back to you.

Participant: [SOAP request]

Journalist: [SOAP response]

Participant: Did you arrive safely home?

Journalist: Sorry, do we know each other?

Participant: Have you developed any ideas for our project?

Journalist: What do you mean?

A.8 Additional Tables and Figures

Provider	# of requests
Work email	9
Twitter	8
LinkedIn	8
Instagram	8
Facebook	7
Gmail	4
Reddit	3
Telekom	2
Pinterest	2
iCloud	2

Table 3: Frequency of how often each IdP was requested in the pilot study. The work email is the provider most frequently requested, and in the current protocol proposal, it is not included.

ID	Sent to			ATI [4]	Reason
	Anne	Amira	Michael		for Failure
CS-1	○	○	○	3.9	
CS-2	○	●	○	6.0	Typosquatting
CS-3	○	●	●	5.6	Typosquatting
CS-4	○	○	○	5.8	
CS-5	○	○	○	5.3	
CS-6	●	●	○	4.4	Typosquatting
CS-7	●	●	●	2.9	Marking
CS-8	○	○	○	2.9	
CS-9	○	○	○	4.4	

Table 4: Overview of the results of the pilot study participants. Four sent data to at least one journalist (marked by ●). The "Reason for Failure" column matches a theme in Section 3.2.6.
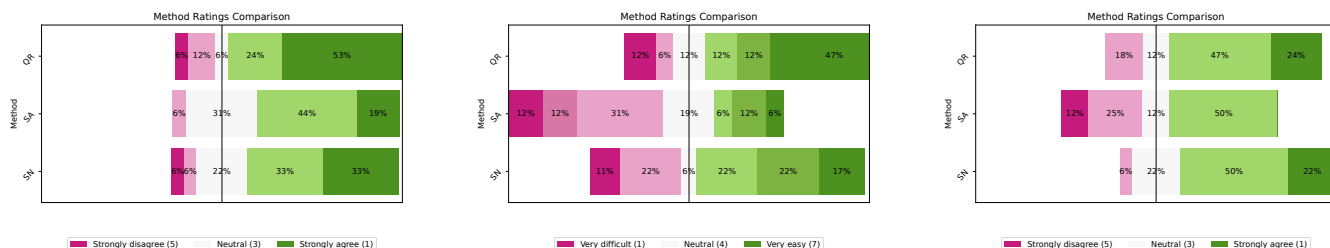
(a) How confident were participants with their decision? (Q8)

(b) Single-ease-question (SEQ) ratings of the methods. (Q9)

(c) How much do the participants trust the method? (Q7)

Figure 2: The ratings of the methods by the participants of the lab studies. The "n"s differ slightly because not each participant used all the methods. "SN" is short for "safety number" and "SA" is short for "social authentication".
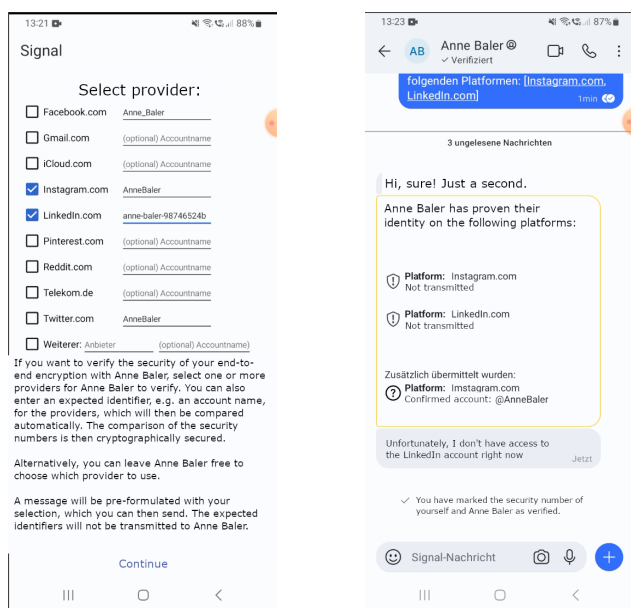
(a) The participant requested two proofs from two providers and filled in the identities.

(b) The response was incorrect due to a typo in one IdP and another identifier not being transmitted; however, the participant did not notice the typo and incorrectly marked Anne as verified.

Figure 3: Translated screenshots of the SOAP request flow from P10 (pilot study).
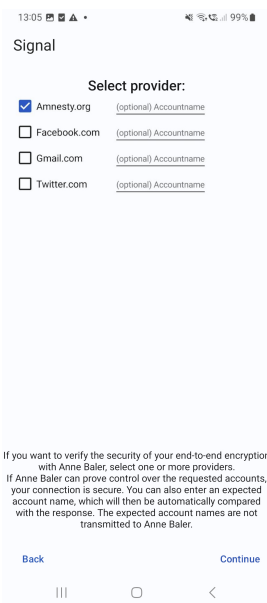
(a) A user can request proof for an account on an IdP without any identity.
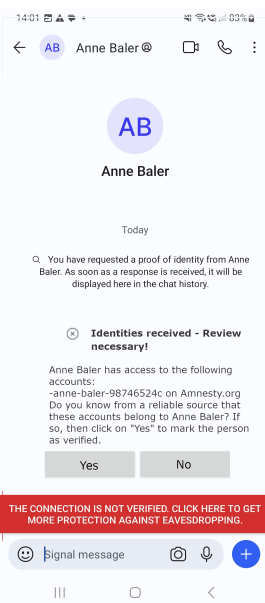
(b) As it cannot be automatically decided whether the identifier is correct, the user must make a decision.
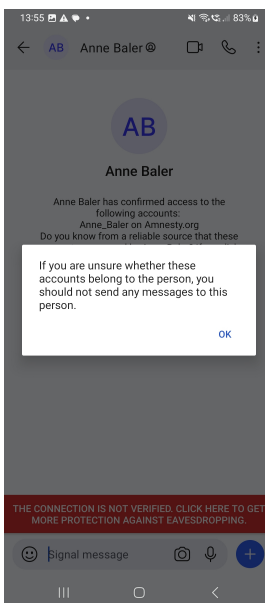
(c) In combination with the decision, the user is informed about possible consequences.
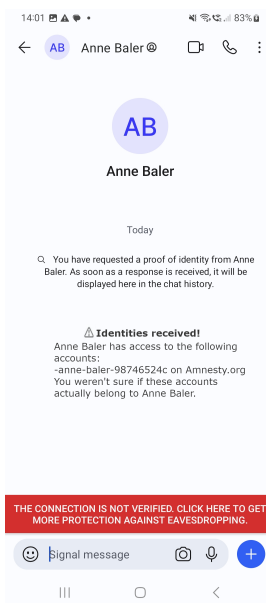
(d) And later reminded what they decided.

Figure 4: Translated screenshots of a SOAP (lab study version) request flow without identifiers.
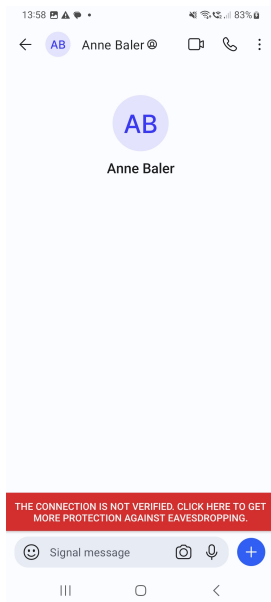
(a) The screenshot shows what a fresh chat looks like if the person was previously added as a contact. The red button nudged the participant to click on it and find the ceremonies.
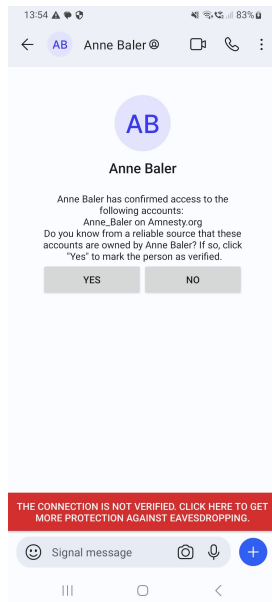
(b) If the participant was in the pre-registration condition, they saw a non-requested SOAP answer.
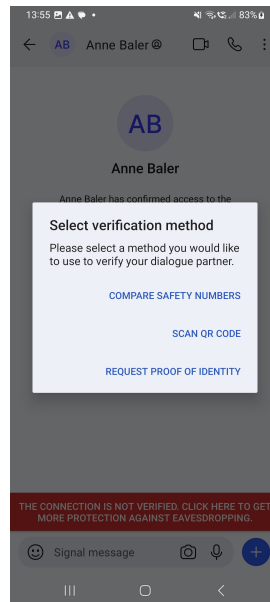
(c) The screenshot shows the menu that led to the ceremonies. Clicking on the QR code and safety number opened the existing site, just slightly modified (see.5d).
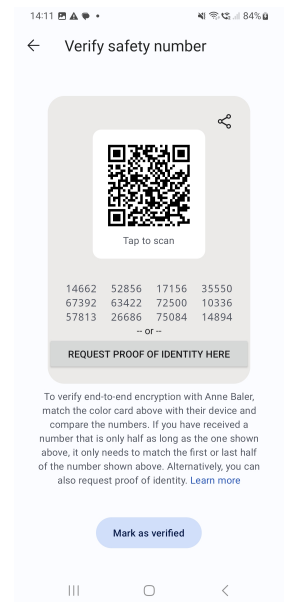
(d) The screenshot shows the menu page where the user can see the safety number, the QR code, and request SOAP. The information text was slightly adapted so that users knew which part of the safety number to compare, and the prompt for social authentication was added.
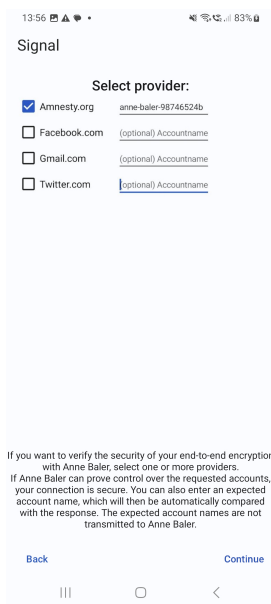
(e) The screenshot shows how a SOAP request code could be made. Selecting an IdP and an identifier. By clicking on next, a SOAP request is sent. The expected identifier is not sent to the recipient.
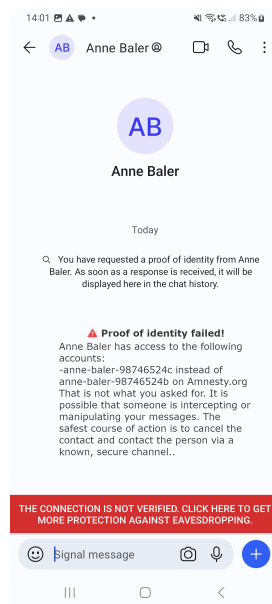
(f) The SOAP request is sent automatically, and the user is informed in the chat. When the user receives a SOAP response, it is parsed, and compared against the expected identifiers and IdPs. This SOAP response is incorrect, as the identifier is not as expected. The user is informed of that.
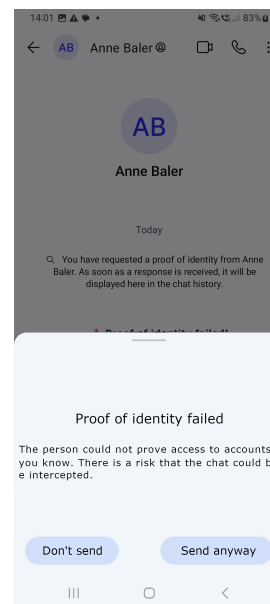
(g) If a user wants to send a message to a contact where a SOAP response was incorrect, the app warns the user, similar to when the safety number of a verified contact changes.
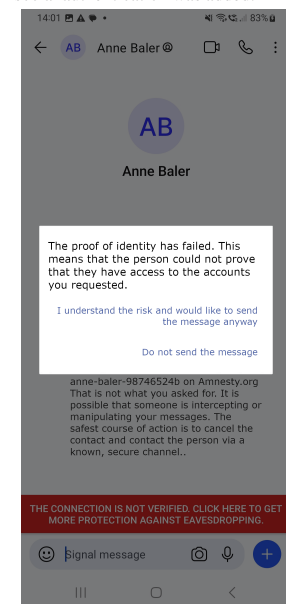
(h) To send a message, the user has to click two additional times.

Figure 5: Translated screenshots of an example SOAP (lab study version) request flow with expected identifiers filled in.

How Entertainment Journalists Manage Online Hate and Harassment

Noel Warford
Oberlin College

Nicholas Farber
University of Maryland

Michelle L. Mazurek
University of Maryland

Abstract

While most prior literature on journalists and digital safety focuses on political journalists, entertainment journalists (who cover video games, TV, movies, etc.) also experience severe digital-safety threats in the form of persistent harassment. In the decade since the #GamerGate harassment campaign against video games journalists and developers, entertainment journalists have, by necessity, developed strategies to manage this harassment. However, the impact of harassment and the efficacy of these strategies is understudied. In this work, we interviewed nine entertainment journalists to understand their experiences with online hate and harassment and their strategies for managing it. These journalists see harassment as a difficult and inevitable part of their job; they rely primarily on external support rather than technical solutions or platform affordances. These findings suggest much more support is needed to reduce the individual burden of managing harassment.

1 Introduction

As part of modern digital life, journalists often have public presences on the internet, through both direct publication and social media. However, when journalists report on things that engender a negative reaction in their audience, they may experience harassment as a result. Although significant prior work has discussed the digital-safety needs and practices of journalists when facing nation-state adversaries [16, 27–30], harassment—defined by Citron as "a persistent and repeated course of conduct targeted at a specific person, that is de-

signed to and that causes the person severe emotional distress, and often the fear of physical harm." [13]—has not been studied as much in this specific context within the digital-safety research community. In this work, we study the experiences and practices of what we term "entertainment journalists"—those who write about topics like movies, video games, and sports—when dealing with online harassment. Despite the less sensitive nature of their work, these journalists experience significant harassment online, which can lead to severe consequences.

We hypothesize that entertainment journalists can experience the intersection of two *contextual risk factors* described in Warford et al.'s framework for understanding the digital-safety needs of at-risk users: *prominence* and *marginalization* [51].[1] *Prominence* refers to users "who stand out in a population, because they are well-known publicly or have noticeable attributes;" *marginalization* refers to "[p]ervasive negative treatment or exclusion at a societal level, due to an individual's identity attributes or life experiences." Warford et al. calls for investigation into the intersection of *contextual risk factors*; this work seeks to answer that call.

We chose to study these journalists due to their experiences dealing with harassment, especially since the 2014 #GamerGate campaign. This campaign targeted feminist video games journalists[2] and developers; #GamerGate supporters claimed to be calling attention to ethics issues in the games journalism industry, but relied primarily on misogynist threats to accomplish that aim [1]. #GamerGate represented the start of a long-term shift toward more organized and pervasive harassment, necessitating stronger protective actions from the campaign's targets [4].

Given this context, we hypothesize that many entertainment journalists have already developed adaptive responses to mitigate harassment's harmful impact, especially if they were targeted by #GamerGate or later campaigns. In this study, we sought to understand harassment's impact on these journal-

---

[1]We note that two of the authors of this paper are also authors of [51].
[2]We use "games journalist" as shorthand for "video games journalist" elsewhere in this paper.

ists and how they attempt to mitigate its harmful effects. Our research questions were as follows:

RQ1: How do entertainment journalists mitigate the negative effects of harassment, both proactively and reactively? How and where do these journalists learn these protective strategies? Are they effective?

RQ2: How is this harassment impacted by risk events (e.g., publishing something about #GamerGate, tweeting about a current controversy)? What are the characteristics of these risk events?

RQ3: How does institutional support play a role in these journalists' protective strategies?

RQ4: What *contextual risk factors* [51] do these journalists experience? How do these impact their experience of online harassment? If multiple contextual factors are at play, how do they interact?

We find that journalists experience particularly damaging harassment attacks, due to the combination of the need for a public profile to promote their work (*prominence*) and the increased impact of harassment on journalists who experience *marginalization*, following prior scholarship on harassment and marginalization [10, 47, 54]. Our participants viewed harassment as an inevitable and dangerous part of their profession. Its consequences included severe mental and emotional impact as well as legitimate fears of escalation to the physical world. Rather than use platform affordances or technical solutions, our participants tended to rely on external support, like colleagues, friends, or family, to manage the impact of harassment on their lives. Improving *external* support for entertainment journalists is critical; they should not have to bear the brunt of harassment alone.

2 Related Work

This work builds on prior scholarship on journalists, online hate and harassment, and #GamerGate, described in this section.

Journalists and digital safety. Journalists have particular digital-safety needs due to their profession. Investigative journalists must securely share relevant documents with trusted sources, but without the subject of investigation finding out [16, 30]. For the high-profile Panama Papers investigation, a customized system for collaboration controlled the flow of information on that investigation without leaks until time of publication, across many different journalists, organizations, and countries [30], but not all journalistic efforts receive an equally high level of attention.

Journalists must also protect their sources—a challenging task in the digital age [25, 27, 28, 38, 43]. Journalists often prioritize communication methods that are "most convenient for the source, including the platform on which [the] source is most likely to respond" [28]. This prioritization can lead to a conflict between the critical need to communicate and the secondary priority of preserving the security of this communication, especially when potential consequences of a security breach could be as severe as imprisonment (for the source) or serious reputational harm (for the journalist). Entertainment journalists may also need to protect sources from retribution, such as when covering topics like harassment [22], labor rights [23, 45], and industry sexual misconduct [20].

Journalists and organizational stakeholders (like editors and IT staff) also have different—and sometimes conflicting—priorities [29]. While both groups may share core concerns like protecting sources and preventing reputational harm, sometimes competing goals put these two groups in conflict. Journalists, for example, may prize collaborating through externally-managed cloud services like Google Docs, but organizational stakeholders may worry about those external services being subpoenaed and revealing sensitive information [29, 31]. Journalists will use technical solutions accomplish their digital-safety goals in high-stakes reporting scenarios [16, 30], so long as these solutions are both clearly valuable and usable.

Most prior literature focuses on journalists protecting sources from nation-states or similarly-resourced adversaries. However, entertainment journalists also face direct attacks from groups of otherwise-ordinary individuals. Their attackers use freely available features of the modern internet and social media in order to target and harass these journalists.

Online hate and harassment. Harassment has become more common over recent years, particularly for young adults (ages 18-24) and LGBTQ+ people, due in large part to "unintended applications of widely accessible technologies" [47]. Thomas et al. taxonomize several important features of modern online hate and harassment, categorizing attacks based on the intended audience, the medium, and the capabilities required [47]. Their threat model includes a *target*—used rather than "victim" in order to empower those who face abuse—and an *attacker*, whose goal is to "emotionally harm or coercively control the target, irrespective of other side effects."

Many other scholars have examined the particular impact of harassment in relationship to experiences of marginalization. Chadha et al. [10] describe how women employ a variety of strategies for dealing with harassment, which is often sexualized, that included normalization and self-censorship. Wei et al. [54] interviewed experts who provide advice for people facing harassment and found that generalizing advice for harassment is difficult, as it relies on the particular type of threat the target faces. For people experiencing marginalization, these experts often added *extra* advice on top of pre-existing best practices, creating an unfair burden for those who face the most severe harassment. In other countries, especially outside

of the Western cultural sphere, participants perceived harassment that damages one's reputation or the reputation of one's family as severely harmful [44]; this reputational damage may escalate into direct physical or sexual violence [42]. The theme of restricting free expression as the only or best option in the face of harassment is a troubling one for researchers who seek to address this problem.

#GamerGate. One specifically salient example of online hate and harassment for this work is the #GamerGate campaign. In August 2014, game developer Zoe Quinn was harassed due to a perception that their game Depression Quest "was lauded with awards, not because of excellent game design and execution, but because it symbolized the gaming world's movement to be more inclusive and progressive" [1]. Quinn's ex-boyfriend alleged an affair between Quinn and games journalist Nathan Grayson, and shared this allegation on the popular imageboard 4chan. This story served as the seed crystal for more severe attacks on Quinn, which then spread to journalists, like feminist games critic Anita Sarkeesian, and game developers, like Brianna Wu, escalating over time to bomb threats and investigation by the FBI [1].

This harassment demonstrates the impact of two *contextual risk factors* identified by Warford et al. [51]—*prominence* and *marginalization*. The "average user" does not usually have to contend with focused harassment from an online group, so targets of #GamerGate who suddenly became *prominent* were often ill-equipped to deal with these threats. This harassment consisted of "extremely offensive and derogatory" [1] language and imagery aimed at women and minorities. Although it is true that not every target of #GamerGate was a woman—for example, Christopher Kushner, founder of 4chan, experienced targeted harassment when banning discussion of #GamerGate on the platform [24]—many attacks relied on existing prejudices against women and marginalized groups.

Structures embedded into social media platforms enable harassment. Trice and Potts show how #GamerGate activists were able to "turn the Twitter experience into an inescapable GamerGate experience" [50]. Since targets were forced to read hateful messages, targets then had to choose between either suffering extreme online abuse or leaving the platform entirely. Massanari describes how the features of Reddit's platform policies—such as an inability to systematically report hateful content and the structure of the platform's homepage—might enable "toxic technocultures" [26]. Chandrasekharan et al. [11] show that banning certain subreddits devoted to hate speech did reduce the level of hate speech on the platform overall, but some users may have migrated to other platforms that were willing to host that content. Online hate and harassment is therefore a structural problem that requires structural solutions.

#GamerGate has also been linked to larger cultural trends in online life. Feminist scholars link the rise of #GamerGate to concerns about feminism playing a greater role in discussion about videogames in the 2010s [18, 19, 32, 49]—although this is not a new topic in feminist scholarship [3, 15, 41, 55], it ironically rose to greater prominence in cultural conversations around digital games in part *due* to #GamerGate [21]. Outside of the sphere of video games, Bezio places #GamerGate as a precursor to the modern "alt-right" movement, especially through the controversy's support by Milo Yiannopoulos and Breitbart [4].

Although the platforms on which this harassment took place have adapted their policies over the intervening ten years, targeted harassment is still an ongoing problem. We use #GamerGate and its influence on the landscape of online harassment to frame our work on entertainment journalists, many of whom were targets themselves during the height of that campaign.

3 Methods

We designed and conducted an interview study to answer our research questions, as described in this section.

3.1 Recruitment

We recruited participants via email and Twitter[3] direct message. We sent messages to candidates who met the following criteria:

- Current or former professional journalist covering media and popular culture (film, television, video games, music, sports, etc.)

- Can work either independently (e.g., YouTube channel, blog, Substack) or for a publishing outlet (e.g., Vice, IGN, Sports Illustrated).

First, we found journalists' contact information through video game websites, given the specific context of #GamerGate. By using prior knowledge and searching for news articles on popular video games, we made a systematic list of websites that publish journalism focused on the video game industry. We also invited interview participants to ask their colleagues if they would be interested in participating. This snowball sampling [34] was essential in getting more journalists to speak with us, given the sensitive nature of the project. As a result, we also interviewed journalists who covered sports, as recommended by our participants, who described sports journalists as common targets for harassment.

Our participants were experienced in the entertainment journalism industry and worked at a variety of outlets. One participant had less than 2 years of experience, one had 5-10 years, and the rest had been working for 11 or more years in the industry. All but one participant had experience at major outlets

---

[3]At the time of recruitment, the social media platform X was still called Twitter, and many of our participants continued to call it Twitter. Rather than use both, we choose to use Twitter here for simplicity.

with many employees, which were often subsidiaries of large media companies with multiple publications, although several participants were freelance at the time of their interviews. Five covered video games (including some who also cover tech culture, tabletop gaming, and other related topics) and four covered sports (including both fan-facing and business-facing coverage). We opt not to present individual participant demographics in detail in order to prevent deanonymization, given these journalists' public presences.

3.2 Protocol

We conducted nine interviews via video conference between April and August 2023. All participants consented to have the interviews audio-recorded. As an extra step to protect participant privacy, we chose not use the audio transcription service built into our video conference software, as this requires sending the recording to a third party with no particular privacy guarantees. Instead, we used WhisperX [2], an open-source speech recognition model which transcribes long-form audio while also identifying distinct speakers.[4] This software was run on a University of Maryland (UMD) computing cluster, ensuring data was only accessible by project researchers and UMD system administrators. The protocol was approved by the UMD IRB.

Our interview protocol covered the following areas. The full protocol can be found in Appendix A.

- **Consent procedures**: Participants were shown the consent form, emphasizing our policies regarding recording, transcription, and anonymity (described below).

- **Warmup and career overview:** We asked about the background of the participant: years worked in the industry, areas of specialty, average readership, size of following on social media, etc.

- **Specific risk events:** We asked about the participant's experience with harassment: what factors seemed to influence their experience of harassment, how common or uncommon this was, and how they managed it. We opted to let participants define harassment for themselves, even as we use Citron's definition in our analysis [13].

- **Advice, given and received:** We asked about where and how participants sought (or gave) advice about dealing with harassment in their networks.

- **Debrief:** We asked participants to describe any relevant personal characteristics that may have impacted their experience, like race, gender, or sexual orientation. We decided *not* to report these systematically, in order to prevent participant deanonymization.

---

[4] https://github.com/m-bain/whisperX

Online hate and harassment can be intensely emotionally challenging. We developed our protocol using a trauma-informed lens, using the following definition of trauma from the Substance Abuse and Mental Health Services Administration (SAMHSA):

> Individual trauma results from an event, series of events, or set of circumstances that is experienced by an individual as physically or emotionally harmful or life threatening and that has lasting adverse effects on the individual's functioning and mental, physical, social, emotional, or spiritual well-being [46].

Chen et al.'s trauma-informed computing framework for computing [12] and Wong's guidelines for trauma-informed care in qualitative research [56] both provide recommendations for conducting trauma-informed research. We adapted these recommendations into the below guidelines for this project:

- We conducted interviews online so that participants could talk to us wherever they were most comfortable.

- We conducted warmup and debrief sessions to establish rapport and regularly checked on participant well-being during the interview.

- We were transparent about the goals of this work.

- We prefaced sections of the interview which discussed difficult topics to give participants a chance to prepare their response.

Anonymity. Harassment events are highly public, so we took extra care in reporting on these events by not just removing names but other identifying details. After each interview, we asked participants if they wanted us to remove any details to protect their privacy and avoid inciting future harassment.

Reflexive thematic analysis. We use *reflexive thematic analysis* in this work to understand our data. First described by Braun and Clarke in 2006 [5], we engage with their most recent perspectives on this methodology [6, 7].

Reflexive thematic analysis posits that rather than excavating ideal truth from one's data, researchers instead create meaning as an active, creative process through the work of interpretation. This interpretation is deliberately situated in the researcher's inherent subjectivity, which is an important part of the process, rather than a bias that should be removed. Rigor is ensured, therefore, by describing the process and situation of the researcher in relation to the work that they are doing – the practice of reflecting on how one's assumptions and process impact the research is *reflexivity*.

Coding and theme development was conducted collaboratively by two researchers, guided by a combination of a

deductive approach—using the at-risk user framework developed by Warford et al. [51] to understand the data—and an *inductive* approach—using the data themselves as a way to understand our participants' experiences. The first two authors independently coded two interviews, discussed discrepancies, and resolved them together; the first author then coded the remaining interviews, and the second author verified these codes by discussing with the first. These authors then created themes together and discussed them with the third author.

Reflexive thematic analysis also exists on a spectrum between *semantic* meaning—focusing on the explicit content of the data—and *latent* meaning—focusing on the implied meaning of the data. On this spectrum, our analysis process tended towards the semantic, focusing on concrete practices and attitudes. However, we analyzed some latent meanings in relationship to how our participants made sense of their experiences of online hate and harassment.

While we bring a particular set of knowledge as computer security researchers, we do not have expertise on our participants' experiences. Rather, the goal of this work is to find the most productive union of the two – by combining the expertise of the digital safety research community *and* the experiences and values of our participants, we can reach the most effective solutions for the unique issues they face.

3.3 Limitations

We do not claim our results are representative of a general population, or even of all entertainment journalists, following appropriate guidelines for reflexive thematic analysis [9]. Instead, the goals of this study were to identify key themes related to our research questions and situate these themes in our participants' contexts and experiences. Accordingly, we did not seek *data saturation*, as that implies an approach contrary to the methods and goals of reflexive thematic analysis [8].

Since harassment is a highly-sensitive and often-traumatic subject, some potential participants with valuable perspectives may have declined to speak with us to avoid further psychological harm. In particular, most of our participants were cisgender men, or present as such online in a way that shields them from gender-based harassment. Many of our participants told us that women, people of color, and trans people experienced even greater levels of harassment than themselves (see Section 4.3). That perspective is largely missing from this work due to this challenge.

Our recruitment process (Section 3.1) also relied on public contact information via Twitter, outlet websites, and other public sources. However, some people who have experienced extreme harassment may deliberately hide their contact information to prevent further harm. Although some of our participants did experience severe harassment, we likely did not capture the full breadth of harassment experiences.

4 The context of harassment

We present our results in two sections. In this section, we describe our participants' understanding of **the context of harassment** across three core themes. In Section 5, we describe how participants choose **strategies for dealing with harassment.** We lightly edited participant quotes by removing some (but not all) filler words or repeated phrases, aiming to capture the tone and style of how our participants speak while still retaining clarity.

4.1 The inevitable price of admission

Our participants commonly described social media——most often, but not exclusively, Twitter——as essential for their jobs. They use social media to contact sources, advertise their work, and receive tips—all critical elements of their profession. However, a public presence on social media simultaneously exposes them to targeted harassment that is impossible to completely avoid. All of our participants had experienced harassment, albeit at different frequencies and levels of intensity; the most experienced participants often had a resigned attitude toward this problem, characterizing it as an inevitable feature of online life.

Participants referenced several platform features as making harassment particularly harrowing ,such as confusion over the effectiveness of moderation tools, shifts in Twitter after Elon Musk's purchase, and the general distortion of reality they experienced on social media. Some participants described looking for other platforms, like replacements to Twitter such as Mastodon or Blue Sky. Others had considered moving to more direct-to-subscriber business models—such as a paid Substack newsletter, a Patreon account, or founding new organizations like Aftermath, the subscriber-supported, worker-owned project from former Kotaku writers [36]—to reduce their need to be active on traditional social media.

Harassment is ephemeral. Although every participant was familiar with harassment, many characterized it as ephemeral, temporary, or impersonal. Many participants believed harassers will inevitably find a new target, particularly if you do not engage (further details on non-engagement as a strategy can be found in Section 5.1). P07 told a story about tweeting that it was important for more women to be involved in sports writing, describing the outcome as follows:

> Some right wing troll account that doesn't cover sports amplified it for whatever reason. And so for like three days I got people on Twitter yelling at me that I was anti-man or that I supported mediocrity in sports writing or four-letter epithets and everything. And just because I wasn't used to it, it really hurt my feelings. And to the point where like my colleagues were like, just give me your phone. And like, don't look at it. Or we're taking away your privileges here

for a little bit. And then, you know, in three days it goes away. Because the people yelling at me don't know who I am. They don't know what my beat is. And it goes away.

Here, P07 also describes a common strategy for managing harassment—having a colleague or trusted other take away your device so you can step away from the online world. Further detail on this strategy can be found in Section 5.3.

Harassment is dangerous. Some participants experienced dramatic escalations of harassment beyond abusive social media comments and hateful emails. P06 called to mind the E3 leak——an incident where the Entertainment Software Association leaked the addresses and contact information of hundreds of journalists attending the video game industry's largest trade show [33]—which exposed his home address. Describing the impact of this experience, P06 said:

> I've had people like paste, like in an email or a DM, just my address. And like, that's it. Like, they don't say anything. [...] They're just able to post that they like, "Hey, we know where you live." Um, that's just information that's out there. [...] It's warm out. It's summer. I don't close every window every night. It's nice to have air going, but you can't help sometimes but wonder. It's like, all right, well, you know, people out there have my address. All it takes is one person to [...] have the wrong idea. And it's like, oh, because I didn't lock the screen door last night, someone can just come into the house.

Participants with children especially feared exposure of their address. As P06 describes above, direct, physical danger could result from having an exposed address due to this leak. Other top-of-mind escalations were receiving upsetting mail or being swatted.[5]

Dangerous outcomes of harassment may also be more likely or more severe for women, people of color, and trans people. Although none of our participants described an escalation that they perceived as related to their identity, many participants believed that harassment is worse for people who experience marginalization.

Harassment has a persistent emotional toll. All participants described harassment as having at minimum a moderate toll on their mental and emotional well-being. For some participants, it was a constant reminder of a group of people who would constantly seek to denigrate them, at times leading to insecurity, as P05 describes:

> There is sort of like a seed planted of just like insecurity. Just like if I voice my opinion online, will

---

[5]Swatting is a common colloquial term for a *false reporting* [47] attack where an attacker will call in a spurious police report in order to get a SWAT team to descend on the home of their target.

people get so mad at me that they're just gonna cuss out my entire existence?

Some participants developed complex post-traumatic stress disorder (PTSD) or severe anxiety based on their experiences. P02 describes the #GamerGate campaign as having this type of severe impact:

> I have complex PTSD. I'm hypervigilant because it was a legitimate, long-time, traumatizing event. So in a really sad way, that hypervigilance is super helpful to anyone who is doing any of this stuff.

Here, P02 also points to the adaptive elements of this condition, discussed further in Section 5.2.

Tensions between characterizations of harassment. The three characteristics of harassment outlined above demonstrate a core tension: how can harassment be ephemeral on the one hand, but dangerous and emotionally draining on the other? These can all simultaneously be true. A specific harassment event, like an insulting tweet or a threatening email, may indeed be a one-time event, and social media furor will eventually pass. Despite this, the *pattern* of these individually-ephemeral events leads to genuine fear of physical harm and attendant psychological consequences.

Where some participants emphasized a "just a name on a screen" (P07) mentality, focusing on social media's inherent disconnection from reality, some instead emphasized that taking action was indeed important. P02, for example, contradicted the common advice to ignore harassment:

> Like you have to take steps. You might not want to engage, but you do have to take a step back or make some kind of statement. But just ignoring it does not help and will often make it significantly worse. Once that fire is going, you have to at least start digging trenches around it so it doesn't spread. So when people start to say, you can just let go and it will flare down, no. You have to do something. Even if it's just getting away from it and taking that step for your own safety. The don't poke the bears thing, yeah, I don't agree with that either. [. . .] I don't think there's a way to ignore the trolls and then they'll go away. We've proven that's not the case.

While P02 still mentions not engaging, here referring to arguing with or making fun of the harasser, he makes clear that *some* kind of action is still required. Shortly thereafter, P02 later vividly compared harassers to toddlers:

> It feels like dealing with a toddler who is trying to get their parents' attention any way they can. And it might, you know, start really innocent and sweet, and then if they're still completely ignored,

---

they might start pushing over lamps. And this is the version of pushing over lamps. And then once it gets to that point, the people who just want to push over lamps, like, yeah, f***ing game on.

4.2 Hurt people hurting people

Based on their experiences, our participants held very strong beliefs about who, exactly, was doing the harassment. Harassers were characterized as "angry", "lonely", "cowards", and generally poorly-adjusted. Some participants pointed out harassers might be seeking some sort of emotional need, often as simple as getting attention or having their voice heard, which occasionally engendered some compassion. P05 said, "There are a lot of instances where like I want to engage with these accounts and just say like "Are you okay?" Because these responses are not like coming from a person who's in the right mindset." It was not clear how these participants reconciled this more sympathetic framing with the real harms that can result from harassment.

Some participants made reference to fans of particular entertainment franchises or sports teams as common sources of harassment. P05, who covers college sports, described how making a small mistake about a team or incorrectly predicting the outcome of a game could lead to harassment far out of proportion to the perceived slight. P03 described the Star Wars and Game of Thrones franchises as attracting particular harassment, especially in relation to "spoiler-phobic culture."

4.3 White cisgender maleness is a shield

Participants who presented as White, cisgender and male on the internet emphasized that this presentation benefited them by protecting them from worse harassment. P08 expressed distress at this phenomenon and described how harassment has driven people without this shield out of his industry:

> The thing that bothers me the most, though, is that, like, I am a privileged cis white male and I have a lot more tools at my disposal that makes it easier for me to, when needed, kind of stand in front of this wind that blows occasionally. And a lot of women and a lot of people of color and a lot of folks less advantaged than me have not been able to do that. And they have had to, pick up and leave their lives and give real aid to those closest to them.

> And some of those voices are the voices that brought me to this career path. There are literally people that brought me into this field who are not here because they weren't able to kind of weather this storm. And it hurts to know that their voices have been silenced and that they are no longer doing this work because of what happened to them. And that hurts at a very intrinsic level because I

value this work so much and I miss them so much. So it lessens my own work not to have their work here next to me being read as well.

Presentation is a key element here. P07 mentioned that, despite having a stereotypically White name and lighter skin, he is Latino, which changed how people spoke to him:

> I'm half Brazilian, like literally whole Brazilian citizenship—mother was from São Paulo, have a Brazilian flag in my account. And so occasionally Latino things come up or about Brazil, I'll talk about those things, but because of my name and because a lot of Americans don't realize that Brazilians can be white too, or white-ish, [they] will feel very comfortable saying something pretty anti-Mexican to me before realizing, oh, he's an immigrant kid too.

P03 mentioned the same phenomenon in terms of their gender identity——despite identifying as genderqueer and using both "he" and "they" pronouns, they were often perceived as a cisgender man. As they say:

> A lot of the time, by being White and presenting as a White guy, I think a lot of time people will take me a little more seriously or be a little less cruel to me than they might otherwise.

Neither P03 nor P07 described intentionally presenting as White or male in order to avoid harassment. Instead, they referred to this phenomenon as passively protective—harassers did not use their identity as a way to attack them, since they were not obviously part of a commonly-marginalized group.

In contrast, our participants who presented as people of color described racism as having a particular impact on their experience of harassment. P05 described needing to pre-emptively mute negative words relating to his identity as a Mexican-American child of immigrants, which required him to think through the worst names someone could call him:

> I can think of any hateful words to describe people, so like the way they antagonize me and like my people, I will add those to the list of muted words so that I don't see them randomly when I'm tweeting about sports or tweeting about like current events.

Even though muting hateful words protected P05 from future bad actors, the task required a significant amount of emotional resilience, as he goes on to discuss:

> It requires me to actually sit down. And at least for a little bit, when I'm in a good, like mental space to come up with as many ways to, you know, to insult me. And, you know, that's no fun either but at least I'm in like a good place to come up with all those as opposed to when I'm feeling down or when I'm currently being harassed.

5 Strategies for dealing with harassment

Now, we discuss how our participants mitigate the effects of harassment and why they chose these methods.

5.1 "Just ignore it": the best worst option

Many of the techniques our participants used to protect themselves from harassment involved some variation on simply ignoring it. As described above, harassment is ephemeral, yet dangerous and emotionally taxing. Nevertheless, most of our participants believed that rather than engage with harassers, it was better to try to ignore them and move on. Many participants stated that any kind of engagement, especially trying to call out harassers' bad behavior, would encourage rather than chastise the harassers—in P06's words, "The moment you start talking about your harassment, you are going to get harassed. [...] That's just fuel for people because they're noticing like, oh, it's getting to them." P03 also describes this in terms of their perception of the harasser's goal:

> These are people just trying to get a rise out of me. And if I give them that, then they get what they're seeking for. And if I ignore them, then they don't.

Participants often described this approach with considerable ambivalence. Although several participants perceive ignoring harassment as the best response most of the time, they did not want to diminish the impact of harassment on others, especially women, trans people, and people of color. Neither ignoring the problem nor confronting harassers seemed to be good solutions; ignoring the problem could feel like tacit endorsement, but confronting it could lead to greater harassment. P06 notes this, particularly when observing others' behavior:

> I have trouble squaring [ignoring harassment] [...] I don't consider that to mean I'm endorsing, like just letting harassment happen. But I do sometimes see amongst people like, they're getting harassed, and their response to it is to like, quote tweet a harasser, and be aggressive with them. Harassment tends to beget more harassment. And so unfortunately, it's like you have a person who is hurting and being harassed. And then, of course, what they're going to want to do is punch back because the platforms are not built in a way to handle this. That is the form of recourse that some people have is just to get angry in response and that frequently seems to just agitate, you know, that's what [harassers] are looking for. And so I don't like that my solution is essentially to just ignore it.

P08 also expressed the impossibility of ignoring harassment that reached a certain level of severity, calling to mind the earlier theme *harassment is dangerous*:

> At the same time that you can't engage, you do kind of need to keep your head on a swivel so that you're aware of what's going to be outside your door when you open it.

Blocking vs. muting. A key technical affordance of Twitter[6] is providing ways to block or mute users. Blocking an account means that account cannot follow you, see what you are saying, or tag you in their own tweets; it also prevents you from seeing the account. Muting, on the other hand, simply prevents an account from showing up on your feed-–—-that user can still reply and see your account, but their activity will not be visible to you.

The primary difference between these two approaches, which have fairly similar outcomes from the perspective of the target of harassment, is that blocking is observable by the blocked user, whereas muting is silent. According to our participants, harassers often perceived being blocked as a badge of honor-—a sign that they had successfully gotten an emotional reaction from you and thus achieved their goal. Muting prevented the harasser from getting what they wanted; the target would not have given the harasser the satisfaction of a strong response. P09 describes this rationale:

> "When people see that they get like, blocked, they see it as like some like, badge of honor, like, look, we, like, defeated this person in some verbal spat, [...] some debate or whatever and [... they] see it as like a badge. So in a weird way, I'd like rather not give you that weird, like, victory in your head that you're right because I blocked you or whatever."

Participants were familiar with block lists-—-automated tools that block huge numbers of accounts that had been collated by others-—-but typically did not use them, due to false positives. Although they might get rid of large amounts of potential harassers at once, the tradeoff of potentially blocking a colleague or friend was seen as not worth the benefit.

Stepping away from the screen. Despite the potential physical danger of harassment described above, our participants often expressed that most harassment has little-to-no relationship with the offline world. When asked about what advice he would give other colleagues, P07 emphasized telling others that "it's important to remember that what you see on TweetDeck is not real life." Putting away your devices and connecting with loved ones in the real world were often crucial strategies for dealing with the emotional impact of harassment. P05 described the following when advising other members of his team, echoing his own strategy to step away during high-harassment events:

---

[6]In this context, we focus on Twitter because our participants did. Other platforms have various blocking affordances that differ from what is described here.

You can scroll through Twitter or Instagram for so long, and then your brain just gets fried. So go outside, go take a break. It might seem counterintuitive for the manager of a social media team to tell you, but like, it's really important to just like log off, tune out and sort of refresh your brain in that way.

This distinction, however, remains difficult in the context of real physical threats. P06, who had experienced a variety of high-harassment events, still characterized harassment as being "99.9% online," despite purchasing physical security after his address was leaked to the public. This apparent contradiction suggests a pattern of low-probability, high-harm events when harassment escalates past insulting comments online.

Look for good-faith actors. A few participants described looking for good-faith actors amongst their social media replies. If someone unskillfully but honestly engaged with the participant's argument, rather than attacking their identity or character, the participant might engage with that person in turn rather than block them. This required energy that other participants were not willing to spend. P04 describes doing this with a surprising sense of playfulness:

> If someone's a little mean but comes in sort of wanting to have a conversation, sometimes I'll send one or two Twitter replies. [...] If someone takes the time to find my email address and sends, like, a mean email, I will sometimes get a little cheeky and be like, "thanks for reading" or, you know, "glad you liked it." Or one time I said, "so does that mean you won't be RSVPing to my birthday party?" You know, stuff like that, depending on the kind of mood that I'm in. If someone took the time to email me.

5.2 Constant vigilance

Potential harassment had a persistent impact on how our participants chose to use social media and how they approached publishing their work. Participants explicitly connected writing about sensitive topics, particularly when speaking with a politically left orientation, and harassment. In some cases, this led to a chilling effect; deliberate self-censorship was sometimes seen as necessary to protect themselves, even about issues important to the participants. P06 describes this effect with a tone of resignation, framing this decision in terms of protecting his family:

> [I have] not really weighed into certain topics to the degree that I might've done in the past, because [...] I have a, broader obligation to think about, which is my family. And there are younger people with more energy than me to sling those arrows and to take them these days.

When describing the impact of harassment on his behavior, P02 described himself as "hypervigilant," (Section 4.1) implying a watchfulness that went over and above what was needed to protect himself due to his prior experiences with severe harassment. He goes on to describe his strategy around *any* social media activity as follows:

> It's high stakes, low risk. It's very rare for a tweet or any message to blow up in a negative way. But if it does, everything's in play on all aspects of your life or the company [which hired you to run their social media]. So when you think about it that way, low risk, high stakes, it's like you kind of do need to bring that care to every single time.

Writing about issues of particular sensitivity would often lead to preparation for potentially being the target of harassment. As an example, when P01's outlet was preparing to cover a game that they anticipated would draw a lot of hateful commentary, they decided to turn off comments in advance and make sure that any journalists who covered the game would be prepared to receive harassment after an article was published. P01's strategy was the most concrete—many participants referenced simply being mentally prepared for harassment when covering a sensitive topic.

Other sensitive topics included the war in Ukraine, violence in video games, anti-racism, feminism, and trans rights, succinctly summarized by P04 as "any issue that's in the culture war at any given time." Participants took varying levels of care when reporting on these topics, depending on their perception of the likelihood and severity of potential harassment. P02 describes a strategy commonly seen amongst colleagues:

> You pay attention to what the big topics are. Right now, for whatever reason, trans athletes, we know that's going to be a big one. Anything having to do with Pride, we know that's going to be a big one. You just kind of keep a list of like things that everyone is talking about in a positive way and things that might be controversial. And most people I know who do this kind of work keep that list in their head. So do I. you just kind of get a sense for it.

Participants with children were particularly mindful of what they shared on social media. In addition to the above self-censorship, P08 took active care to delete images of his children from Instagram when Twitter, in his view, started to decline:

> Preemptively, as Twitter began to melt down, I went to my Instagram and I deleted every image of my children out of my Instagram to kind of sanitize that and make that a place where I could land professionally, if need be.

5.3 External support is critical

Participants relied on their social networks and employers in order to support them while experiencing harassment. This was expressed in two main ways: concrete support from employers and emotional support from friends and colleagues.

Concrete support. Our participants often relied on their news organizations as a primary vector for concrete solutions. These companies provided a variety of resources, often paying for protective tools and services. For more severe incidents, these organizations also provided legal support or paid for cameras to protect their journalists' homes. For P08, going independent (as opposed to working for a larger organization) would be terrifying, as not having "DC lawyers" on his side would make facing harassment vastly more challenging. As he described, "That's a good feeling to go to bed with at night, the next morning, no matter what happened at work that day. to know that there are some angry, smart motherf***ers with law degrees in my corner."

Not every organization always had these policies——P08 described how his publication did not have a concrete harassment mitigation policy until the #GamerGate controversy targeted him and his colleagues. However, this led to improvements for *other* publications in the same parent news organization who learned from P08 and his colleagues' experiences:

> Our expertise at [news organization] was actually crucial in supporting some of our other verticals[7] as they entered the 2016 election season, for instance.

Participants often mentioned data deletion services like DeleteMe[8]——companies that, as a service, will search the internet for one's personal details and get them deleted—which were sometimes paid for by their employer, and other times paid for by the participants themselves.

Colleagues and management also offered support mechanisms, both formal and informal. P01 and P03 both described a Slack channel where their colleagues could share stories, provide information, and seek support when facing harassment. Several participants made reference to giving colleagues their phones to perform triage in the face of severe harassment——this meant that the person who was suffering harassment did not have to deal with blocking, muting, and otherwise managing the incident; they could instead take some time for themselves to recover and let the incident pass.

Two participants mentioned local police, although in contradictory terms. P08 described educating his local police department about swatting (defined in Section 4.1). He remained in regular contact with his police department to ensure the threat was accurately understood. P06, on the other hand,

described his police department as oblivious to this threat, despite repeated attempts to educate them. No other participants mentioned police helping to handle or track down threats.

Emotional support. Even more than the above concrete support strategies, emotional support from colleagues was essential. P03 describes their colleagues setting an example for how to respond to harassment:

> I remember when I came up on the college football team, 'cause that was, like, the work environment that really cemented a lot of how I approach things in me. You know, it was my first full-time job in journalism and [. . .] I was the youngest person on the team, you know?

> So it's like one of those situations where everyone else there, you're kind of looking up to them to set an example. And a lot of those guys were, really funny, smart, like the smartest, funniest people I knew. They were all kind of brash Southern guys who like were aligned with me on like morals and stuff, but also were like really into college football.

> And I say this stuff about them being the funniest, smartest guys I know, because then they would still just get so many stupid people saying just, like, heinous things to them on social media in the comments and the way they responded was mostly to laugh.

Our participants often described how simply sharing their experience with colleagues was valuable for keeping themselves healthy in the face of extreme harassment. P09 also referenced seeking out other Black games journalists, because their particular experiences around harassment and race were more specific, and therefore more useful.

> It just kind of became a thing where, like, if I ever had, like, a more specific kind of question about harassment and things that I face, I'd feel more comfortable to ask someone in this space who [...] would be Black and would probably have the same kind of, like, avenue of harassment that I'd face. So I kind of would reach out to them as like point people to be like, "so what can I expect getting into this role?" And I'd like kind of talk to [them] on and off kind of just about like specific harassment stuff. So like we'd kind of, like, be like go-to points for each other for the most part for stuff.

Some participants experienced severe mental health consequences and sought external support from therapists. Sometimes this was helpful, but P04 describes a therapist dismissing his concerns after a period of intense harassment:

> I even went to therapy because I was bothered so much by [harassment]. Therapist saw me once and

---

[7]In this context, a vertical refers to a news site that is dedicated to one particular topic, often covering it in more detail and with more analysis than a generalist news publication.

[8]https://joindeleteme.com/

said he wouldn't see me again because this wasn't actually a mental illness. I was like, "yeah, that's fair." But I was looking for any tips on how to deal with it because I was letting it occupy too much of my brain space.

Support from family and friends was also crucial, especially when stepping away from the screen and thus from the source of the harassment.

5.4 IT hygiene

Our participants' toolkits included a certain amount of basic "IT hygiene" (P04). Some mentioned multi-factor authentication as a useful tool, particularly to prevent a harasser from taking over their account and causing severe reputational harm. P05 describes his approach:

> The one thing that I'm sort of more focused on sort of in conjunction with harassment is just getting my accounts hacked by people who sort of want to troll me or just want to take it a step further. And because of that, I've just gotten into just having my accounts as secure as possible, whether it's like setting up two-factor authentication or physical security keys, so that even if somehow my accounts get hacked, they still won't be able to access stuff and post as me, impersonating as me, so that they ruin my professional and maybe personal life because I've seen instances of that.

P05 then described watching for malicious links and limiting which devices were logged into his accounts. Perhaps surprisingly to the digital-safety research community, only P04 and P05 referenced traditional security advice. For targets of harassment, security advice may need to be more contextual to be useful, as prior work has demonstrated [39, 54].

6 Discussion

In this work we show that entertainment journalists experience severe harassment—while any *individual* insult may be ephemeral, the *pattern* of harassment exacts a severe emotional toll and can escalate to real danger. In response, our participants largely adopted an "ignore it and move on" attitude, choosing to disengage in order to protect themselves. They rely on external support to help manage this, both practically and emotionally. We also investigate the intersection of *prominence* and *marginalization*—harassment directed at entertainment journalists who experience marginalization may target their identity, leading to more severe outcomes. These findings echo the taxonomy of harassment presented in Thomas et al. [47]—our participants experienced *toxic content* (e.g., bullying, threats, sexual harassment), *content leakage* (e.g., revealing personally-identifiable information, doxxing) and

overloading (e.g., forcing the target to triage hundreds of notifications). They were also worried about the possibility of *false reporting* in the form of swatting.

6.1 Targets of harassment often must fend for themselves

Many protective strategies described by our participants emphasized individual action in the face of harassment. From muting hateful words that targeted one's identity (Section 4.3) to blocking users who send harassing comments (Section 5.1), our participants' strategies required taking personal responsibility for managing online harassment. Since the attackers in this scenario are anonymous online mobs, this is inherently unbalanced.

This finding echoes prior work on harassment of *prominent* individuals. Content creators (defined as "social media personalities with large audiences on platforms like Instagram, TikTok, and YouTube" [48]) also characterized harassment as unavoidable. Like our participants, they experience harassment largely in the form of *toxic content* and *overloading*, which generally must be managed individually. Even security experts who understand the severe burdens of managing harassment tend to provide advice that is focused on individual action rather than systemic change [54]. Mandating individual responsibility for dealing with harassment can lead to perpetuating these patterns by dismissing the societal factors that lead to *marginalization*.

In part due to this imbalance, both our participants and other *prominent* individuals rely on distancing behaviors and their social networks to manage this threat, rather than technical solutions; for people who experience *marginalization*, this is even more pronounced (e.g., [48]). This reinforces the findings of Warford et al., who describe distancing behaviors and social strategies as core pillars of at-risk users' response to digital-safety threats [51]. Technical solutions may indeed be useful, but they must be relevant, targeted, and discoverable.

Importantly, in contrast to prior work, we find that our participants often rely on *institutional* external support, in addition to friends and colleagues. We show that news organizations can provide concrete, useful supports like legal and financial assistance to their employees (Section 5.3). Many participants described how having external support increased their peace of mind, both in terms of concrete assistance from their employer and emotional support from their colleagues. This is an important element to consider when developing solutions to prevent or mitigate harassment; tools and techniques that rely solely on individual action rather than leveraging their communities may be missing a key piece of the harassment mitigation process.

6.2 Moving toward collective responsibility

The imbalance of responsibility we discuss in Section 6.1 suggests that a rebalancing is necessary. Our results suggest several potential avenues for improvement.

Community support and mutual aid. One promising avenue of solutions might be mechanisms for explicitly supporting mutual aid among colleagues. Community resources, like shared lists of muted words, could relieve some of the burdens faced by *marginalized* individuals. Muting hateful language, as P05 described (Section 4.3), is a difficult and draining task; if communities could conveniently create and share crowd-sourced lists of muted words, this could alleviate some of that burden. Although it would not solve the entire problem— after all, someone would still need to add words to this list— methods like these could allow individuals to rely on their communities more effectively for support.

P09 additionally described commiseration with other Black writers as supportive, both preemptively and after experiencing harassment (Sections 4.3 and 5.3). This required P09 to reach out individually to trusted others, which again required him taking personal responsibility for this societal problem. It might be useful to create formal social structures that are run by and for people who experience harassment. Especially for independent journalists, having a community of supportive colleagues could provide the emotional benefits described by our participants.

Changes at the platform level. Our participants did not use many platform affordances to mitigate harassment. Even when they did, they described these affordances as often unhelpful or unclear, suggesting platform-level improvements are necessary. For example, social media sites could build in tools to allow users to assign someone else to triage their account. Currently, our participants described handing over their entire phone or social media account; even a well-meaning helper might see something they did not intend under this model. If developed properly, these systems could limit access to certain apps, restrict access to only desired parts of platforms, automatically revoke access after a certain amount of time, or some configurable combination of the above. These systems could be approached with an eye toward mutual aid— users might take shifts or work ad-hoc to help others in their network, using these tools to triage high-harassment events.

High-quality moderation can also help, but that carries its own costs—shifting the labor of dealing with hateful speech from the targets to the invisible-but-indispensable commercial content moderation workers who already act as the first point-of-contact for hate and harassment [40]. Automated hate speech detection and intervention is also a promising area of future work [17, 37, 53], but challenges of accuracy and equity remain [14, 35, 52].

Assisting organizations in supporting their employees. Providing institutions with the knowledge and resources to support their employees would also be helpful. As described by many of our participants, the backing and support of formal institutions was supportive. For example, in Section 5.3, for example, P08 describes sharing strategies for mitigating harassment with political reporters to build capacity for managing harassment across the entire organization. However, not all news organizations already have this institutional knowledge, so researchers could create resources to help these organizations learn how to provide needed protections to their employees.

Support organizations like Tall Poppy[9] and PEN America[10] exist to address these concerns, but our participants did not mention them in the context of support strategies. To date, these organizations have focused on other areas: Tall Poppy has mostly worked with streaming providers like Twitch or Spotify, and PEN America focuses largely on literature and free speech. Both organizations' expertise, however, could be very useful to journalists, both in entertainment and elsewhere. A bidirectional relationship between support organizations and news organizations would therefore be beneficial; this would mean that news organizations do not have to develop new expertise, but can rely on the previous experience of these organizations.

7 Conclusion

Entertainment journalists experience severe harassment online. Although this harassment is pervasive, dangerous, and constant, entertainment journalists see it as the price of admission into their chosen profession, since they need to use social media platforms to promote their work. At present, the technical tools available to these journalists are insufficient; platform affordances do little to prevent the flood of harassment these journalists experience. Many participants found simply ignoring the harassment was the best and only option, rather than engaging with the harassers. As a result, participants relied on external support — colleagues, friends, and family — to ameliorate the negative effects of this experience. Therefore, a greater emphasis on *non-technical* solutions to sociotechnical problems could be of great value, in addition to continued development of technical approaches.

Acknowledgements

---

[9] https://www.tallpoppy.com/
[10] https://pen.org/

References

[1] Sarah A. Aghazadeh, Alison Burns, Jun Chu, Hazel Feigenblatt, Elizabeth Laribee, Lucy Maynard, Amy L. M. Meyers, Jessica L. O'Brien, and Leah Rufus. GamerGate: A case study in online harassment. In *Online Harassment*, pages 179–207. 2018.

[2] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. WhisperX: Time-accurate speech transcription of long-form audio. In *INTERSPEECH*, 2023.

[3] Elizabeth Behm-Morawitz and Dana Mastro. The effects of the sexualization of female video game characters on gender stereotyping and female self-concept. *Sex Roles*, 61(11-12):808–823, 2009.

[4] Kristin MS Bezio. Ctrl-Alt-Del: GamerGate as a precursor to the rise of the alt-right. *Leadership*, 14(5):556–566, 2018.

[5] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006.

[6] Virginia Braun and Victoria Clarke. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, 11(4):589–597, 2019.

[7] Virginia Braun and Victoria Clarke. *Thematic Analysis: A Practical Guide*. SAGE Publications, 2021.

[8] Virginia Braun and Victoria Clarke. To saturate or not to saturate? Questioning data saturation as a useful concept for thematic analysis and sample-size rationales. *Qualitative Research in Sport, Exercise and Health*, 13(2):201–216, 2021.

[9] Kelly Caine. Local standards for sample size at CHI. In *Proc. CHI*, 2016.

[10] Kalyani Chadha, Linda Steiner, Jessica Vitak, and Zahra Ashktorab. Women's responses to online harassment. *International Journal of Communication*, 14:239–257, 2020.

[11] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. You can't stay here: The efficacy of Reddit's 2015 ban examined through hate speech. *PACM HCI*, 1(CSCW):1–22, 2017.

[12] Janet X. Chen, Allison McDonald, Yixin Zou, Emily Tseng, Kevin Roundy, Acar Tamersoy, Florian Schaub, Thomas Ristenpart, and Nicola Dell. Trauma-informed computing: Towards safer technology experiences for all. In *Proc. CHI*, 2022.

[13] Danielle Keats Citron. Addressing cyber harassment: An overview of hate crimes in cyberspace. *Case Western Reserve Journal of Law, Technology and the Internet*, 6, 2015.

[14] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proc. ICWSM*, 2017.

[15] Tracy L Dietz. An examination of violence and gender role portrayals in video games: Implications for gender socialization and aggressive behavior. *Sex Roles*, 38(5/6), 1998.

[16] Kasra EdalatNejad, Wouter Lueks, Julien Pierre Martin, Soline Ledésert, Anne L'Hôte, Bruno Thomas, Laurent Girod, and Carmela Troncoso. DatashareNetwork: A decentralized privacy-preserving search engine for investigative journalists. In *Proc. USENIX Security*, 2020.

[17] Paula Fortuna and Sérgio Nunes. A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51(4):85:30, 2018.

[18] Armanda Gonzalez, Elaine Gomez, Rebeca Orozco, and Samuel Jacobs. Entering the boys' club: An analysis of female representation in game industry, culture, and design. In *Proc. iConference*, 2014.

[19] Kishonna L. Gray, Bertan Buyukozturk, and Zachary G. Hill. Blurring the boundaries: Using GamerGate to examine "real" and symbolic violence against women in contemporary gaming culture. *Sociology Compass*, 11(3):12458, 2017.

[20] Kirsten Grind, Ben Fritz, and Sarah E. Needleman. Activision CEO Bobby Kotick knew for years about sexual-misconduct allegations at videogame giant. *Wall Street Journal*, 2021. https://www.wsj.com/articles/activision-videogames-bobby-kotick-sexual-misconduct-allegations-11637075680.

[21] Matthew Handrahan. IGDA: Gender, GamerGate and the need for action. *GamesIndustry.biz*, 2015. https://www.gamesindustry.biz/articles/2015-04-29-igda-gender-gamergate-and-the-need-for-action.

[22] Patrick Klepek. Nintendo employee 'terminated' after smear campaign over censorship, company denies harassment was factor. *Kotaku*, 2016. https://kotaku.com/nintendo-employee-terminated-after-smear-campaign-over-1768100368.

[23] Patrick Klepek. How restrictive contracts stifle and control creativity in the video game industry. *Vice*, 2023. https://www.vice.com/en/article/g5va43/noncompete-contracts-video-game-industry.

[24] David Kushner. 4chan's overlord Christopher Poole reveals why he walked away. *Rolling Stone*, 2015. https://www.rollingstone.com/culture/culture-features/4chans-overlord-christopher-poole-reveals-why-he-walked-away-93894/.

[25] Paul Lashmar. No more sources?: The impact of Snowden's revelations on journalists and their confidential sources. *Journalism Practice*, 11(6):665–688, 2017.

[26] Adrienne Massanari. #GamerGate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*, 19(3):329–346, 2017.

[27] Susan E. McGregor. Digital security and source protection for journalists. Technical report, Tow Center for Digital Journalism, 2014.

[28] Susan E. McGregor, Polina Charters, Tobin Holliday, and Franziska Roesner. Investigating the computer security practices and needs of journalists. In *Proc. USENIX Security*, 2015.

[29] Susan E. McGregor, Franziska Roesner, and Kelly Caine. Individual versus organizational computer security and privacy concerns in journalism. *PoPETs*, 2016(4):418–435, 2016.

[30] Susan E. McGregor, Elizabeth Anne Watkins, Mahdi Nasrullah Al-Ameen, Kelly Caine, and Franziska Roesner. When the weakest link is strong: Secure collaboration in the case of the Panama Papers. In *Proc. USENIX Security*, 2017.

[31] Susan E. McGregor, Elizabeth Anne Watkins, and Kelly Caine. Would you Slack that?: The impact of security and privacy on cooperative newsroom work. *PACM HCI*, 1(CSCW):1–22, 2017.

[32] Torill Elvira Mortensen. Anger, fear, and games: The long event of #GamerGate. *Games and Culture*, 13(8):787–806, 2018.

[33] Maddy Myers. E3 expo leaks the personal information of over 2,000 journalists. *Kotaku*, 2019. https://kotaku.com/e3-expo-leaks-the-personal-information-of-over-2-000-jo-1836936908.

[34] Mahin Naderifar, Hamideh Goli, and Fereshteh Ghaljaie. Snowball sampling: A purposeful method of sampling in qualitative research. *Strides in Development of Medical Education*, 14(3), 2017.

[35] Debora Nozza. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proc. ACL-IJCNLP*, 2021.

[36] Jay Peters. Former Kotaku writers are launching a new video game site — and they own it this time. *The Verge*, 2023. https://www.theverge.com/2023/11/7/23949269/aftermath-video-games-kotaku-defector.

[37] Flor Miriam Plaza-del-arco, Debora Nozza, and Dirk Hovy. Respectful or toxic? Using zero-shot learning with language models to detect hate speech. In *Proc. WOAH*, 2023.

[38] Julie Posetti. *Protecting Journalism Sources in the Digital Age*. UNESCO Publishing, 2017.

[39] Elissa M. Redmiles, Noel Warford, Amritha Jayanti, Aravind Koneru, Sean Kross, Miraida Morales, Rock Stevens, and Michelle L. Mazurek. A comprehensive quality evaluation of security and privacy advice on the web. In *Proc. USENIX Security*, 2020.

[40] Sarah T. Roberts. Commercial content moderation: Digital laborers' dirty work. *Media Studies Publications*, 2016.

[41] Pam Royse, Joon Lee, Baasanjav Undrahbuyan, Mark Hopson, and Mia Consalvo. Women and games: Technologies of the gendered self. *New Media & Society*, 9(4):555–576, 2007.

[42] Nithya Sambasivan, Amna Batool, Nova Ahmed, Tara Matthews, Kurt Thomas, Laura Sanely Gaytán-Lugo, David Nemer, Elie Bursztein, Elizabeth Churchill, and Sunny Consolvo. "They don't leave us alone anywhere we go": Gender and digital abuse in South Asia. In *Proc. CHI*, 2019.

[43] Charlie Savage and Leslie Kaufman. Phone records of journalists seized by U.S. *The New York Times*, 2013. https://www.nytimes.com/2013/05/14/us/phone-records-of-journalists-of-the-associated-press-seized-by-us.html.

[44] Sarita Schoenebeck, Amna Batool, Giang Do, Sylvia Darling, Gabriel Grill, Daricia Wilkinson, Mehtab Khan, Kentaro Toyama, and Louise Ashwell. Online harassment in majority contexts: Examining harms and remedies across countries. In *Proc. CHI*, 2023.

[45] Jason Schreier. The horrible world of video game crunch. *Kotaku*, 2016. https://kotaku.com/crunch-time-why-game-developers-work-such-insane-hours-1704744577.

[46] Substance Abuse and Mental Health Services Administration. SAMHSA's concept of trauma and guidance for a trauma-informed approach. Technical report, 2014. https://ncsacw.acf.hhs.gov/userfiles/files/SAMHSA_Trauma.pdf.

[47] Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley, Deepak Kumar, Damon McCoy, Sarah Meiklejohn, Thomas Ristenpart, and Gianluca Stringhini. SoK: Hate, harassment, and the changing landscape of online abuse. In *Proc. IEEE S&P*, 2021.

[48] Kurt Thomas, Patrick Gage Kelley, Sunny Consolvo, Patrawat Samermit, and Elie Bursztein. "It's common and a part of being a content creator": Understanding how creators experience and cope with hate and harassment online. In *Proc. CHI*, 2022.

[49] Cherie Todd. Commentary: GamerGate and resistance to the diversification of gaming culture. *Women's Studies Journal*, 29(1):64–67, 2015.

[50] Michael Trice and Liza Potts. Building dark patterns into platforms: How GamerGate perturbed Twitter's user experience. *Present tense*, 6, 2018.

[51] Noel Warford, Tara Matthews, Kaitlyn Yang, Omer Akgul, Sunny Consolvo, Patrick Gage Kelley, Nathan Malkin, Michelle L. Mazurek, Manya Sleeper, and Kurt Thomas. SoK: A framework for unifying at-risk user research. In *Proc. IEEE S&P*, 2022.

[52] Mark Warner, Angelika Strohmayer, Matthew Higgs, and Lynne Coventry. A critical reflection on the use of toxicity detection algorithms in proactive content moderation systems. http://arxiv.org/abs/2401.10629, 2024.

[53] Mark Warner, Angelika Strohmayer, Matthew Higgs, Husnain Rafiq, Liying Yang, and Lynne Coventry. Key to kindness: Reducing toxicity in online discourse through proactive content moderation in a mobile keyboard. http://arxiv.org/abs/2401.10627, 2024.

[54] Miranda Wei, Sunny Consolvo, Patrick Gage Kelley, Tadayoshi Kohno, Franziska Roesner, and Kurt Thomas. "There's so much responsibility on users right now:" expert advice for staying safer from hate and harassment. In *Proc. CHI*, 2023.

[55] Dmitri Williams, Mia Consalvo, Scott Caplan, and Nick Yee. Looking for gender: Gender roles and behaviors among online gamers. *Journal of Communication*, 59(4):700–725, 2009.

[56] Rebecca Wong. Guidelines to incorporate trauma-informed care strategies in qualitative research. 2021. https://www.urban.org/urban-wire/guidelines-incorporate-trauma-informed-care-strategies-qualitative-research.

A Interview protocol

Introduction. Hi, my name is [researcher name], thanks for agreeing to participate in this research - we really appreciate your time.

First, let's quickly go over how this study will work. I will be interviewing you and [I/my colleague] will take notes. I expect the study to take approximately one hour. . One thing I'd like to mention is that the research interview process is somewhat different than the journalistic interview process - rather than seeking pull quotes or particularly interesting stories, we are instead looking for common themes between an entire corpus of interviews, even if those themes seem at first to be mundane.

[Describe everything on the consent form.]

We may cover some sensitive topics during this interview, so if you become uncomfortable at any time during the study and wish to withdraw, please let me know. You are also welcome to skip any questions you do not wish to answer, and you only need to provide as much information as you're comfortable with. Do you have any questions so far?

[Give the participant the link to the consent form.]

This consent form tells you who to contact if you have any problems or want to report any objections. Please print a copy of this consent form for your records.

[point out places the subject needs to mark checkboxes]

We would like to record the audio of this interview with your permission in order to properly represent your statements and point of view. However, recording is optional - if, either now or after the fact, you would like us to not use or delete the recording of this interview, please let me know. I'm also happy to answer questions about how we store and use these recordings. We will also be taking written notes during the interview. Do you give permission for us to audio-record this interview?

[If they agree, the interview was recorded. If not, the interviewer took notes.]

Warmup/Career summary.

1. Can you please describe your career in media journalism?

2. How did you get started?

3. When did you get started?

4. What outlets have you worked at over the course of your career?

5. Do you have a particular specialty, like esports, a particular media property, interviews with creators, opinion pieces, or something else?

6. Can you tell me about something interesting you've been working on recently?

7. Could you tell us a little about your online engagement with readers/viewers?

8. What social media platforms do you use today? About how many followers on each platform did you have? How many readers or viewers do you usually reach?

Questions about Specific Risk Events. In the next part of the study, we are going to ask you about your experiences with harassment. We emphasize that we do not view any harassment as justified, but we also acknowledge that sometimes the amount of harassment one experiences might vary at different times or when writing about certain subjects. For the purposes of this study, we are going to define a "high harassment event" as a short period of time with an unusually high quantity or intensity of harassment.

1. Please describe your experience of online harassment on a day-to-day basis.

2. In the past two years, have you experienced any high harassment events? Please describe it in as much detail as you feel comfortable.

3. What patterns, if any, have you noticed in how and when high harassment events occur? [below prompts if necessary]

   (a) External events - either related to your industry or not

   (b) Publishing articles or social media posts about a certain topic or issue (For example, when I tweet about X, I get a ton of angry DMs)

   (c) Publishing any kind of article or social media post

4. How do you typically respond when you experience online harassment? [the following questions may be asked as needed for each protective strategy]

   (a) Did you take this protective action because you anticipated an increase in harassment, or after the increase in harassment started?

   (b) How effective did you feel [this strategy] was? Why do you feel it is effective/ineffective?

   (c) Do any colleagues or friends or people you know employ [this strategy] Is it effective for them? Why or why not?

   (d) How did you learn about [this strategy]?

5. There are all kinds of strategies people use in situations like this - certain strategies work for some people, but not for others. Are there any harassment protection strategies you have heard about / considered but did not take? [the following questions may be asked as needed for each protective strategy]

   (a) Why or why not? [Prompts follow if the participant has difficulty answering]

   (b) Does [strategy] not work generally? Why?

   (c) Is [strategy] not applicable to your particular situation? Why?

   (d) Does [strategy] have costs or downsides that make it difficult/unrealistic/undesirable to implement? What are those costs or downsides?

   (e) Did you ever employ [strategy] in the past? Why did you stop employing it?

6. Are there any tools or technologies you use to respond to harassment? This can include affordances of various platforms (like blocking an individual on social media) or external tools (like blocklists that can be shared amongst users).

7. Just like strategies, different people use different tools and technology to deal with harassment for different reasons. Are there any tools or technologies for responding to harassment that you know about but do not use?

   (a) Why or why not? [Prompts follow if the participant has difficulty answering]

   (b) Does [tool] not work generally? Why?

   (c) Is [tool] not applicable to your particular situation? Why?

   (d) Does [tool] have costs or downsides that make it difficult/unrealistic/undesirable to implement? What are those costs or downsides?

   (e) Did you ever use [tool] in the past? Why did you stop using it?

8. In as much or as little detail as you like, could you please describe the impact of this experience on your life? You're welcome to discuss either specific events or your general experience of harassment.

   (a) How does it impact you, emotionally?

   (b) How does it impact your career?

   (c) How does it impact your relationships with others?

Advice, Given and Received. Now, I'm going to ask some questions about specific strategies and advice you may have heard of or used when managing harassment.

1. What security advice have you received in the past that's relevant to your experience? [for each item, ask the following if needed]

    (a) If unsure what I mean by security advice: Some examples of security advice might include "use a password manager", "don't respond to harassers", or "log off from social media for a while".

    (b) Did you follow this advice? Why or why not?

    (c) How did you hear about this advice? If no response: could prompt for "on the Internet", "from colleagues", etc.

    (d) In your opinion, how effective is this advice in relation to achieving your security goals?

    (e) How difficult is this advice to follow?

    (f) How time-consuming is it to implement this advice?

    (g) How confident are you that you could implement this advice?

    (h) How disruptive would it be to implement this advice?

2. Do you have any trusted people in your network you can turn to for advice, either on security specifically or in general with regards to responding to harassment?

3. Have you ever given advice to someone in a similar situation as yours? What advice did you give?

    (a) Did they follow this advice, to your knowledge? Why or why not?

    (b) Is [advice] not applicable to their particular situation? Why?

    (c) Does [advice] have costs or downsides that make it difficult/unrealistic/undesirable to implement? What are those costs or downsides?

Closing.

1. For this study, we are choosing not to systematically collect common demographic data, like race, gender, age and ethnicity, in order to help protect participant privacy. However, we acknowledge these factors can have an impact on one's experience of harassment, so I'd like to give you the opportunity now to share any identifiers you are okay with us reporting as part of our analysis. We will, of course, not use your name, but we can also mask or share other factors.

2. Please share any comments, suggestions, or feedback you have about our study.

'Custodian of Online Communities': How Moderator Mutual Support in Communities Help Fight Hate and Harassment Online

Madiha Tabassum
Northeastern University
Boston, MA, USA
m.tabassum@northeastern.edu

Alana Mackey
Wellesley College
Wellesley, MA, USA
am116@wellesley.edu

Ada Lerner
Northeastern University
Boston, MA, USA
a.lerner@northeastern.edu

Abstract

Volunteer moderators play a crucial role in safeguarding online communities, actively combating hate, harassment, and inappropriate content while enforcing community standards. Prior studies have examined moderation tools and practices, moderation challenges, and the emotional labor and burnout of volunteer moderators. However, researchers have yet to delve into the ways moderators support one another in combating hate and harassment within the communities they moderate through participation in meta-communities of moderators. To address this gap, we have conducted a qualitative content analysis of 115 hate and harassment-related threads from r/ModSupport and r/modhelp, two major subreddit forums for moderators for this type of mutual support. Our study reveals that moderators seek assistance on topics ranging from fighting attacks to understanding Reddit policies and rules to just venting their frustration. Other moderators respond to these requests by validating their frustration and challenges, showing emotional support, and providing information and tangible resources to help with their situation. Based on these findings, we share the implications of our work in facilitating platform and peer support for online volunteer moderators on Reddit and similar platforms.

1 Introduction

Social media platforms, such as Reddit and Facebook, have emerged as powerful hubs for individuals seeking and providing support across diverse communities. These platforms facilitate the formation of online spaces where users can connect with like-minded individuals or those facing similar challenges, finding solace and understanding as they share experiences, seek advice, and offer empathy. These digital communities span various topics, including mental health, chronic illnesses, parenting, etc. The immediacy and accessibility of these platforms enable individuals to find support at any time, transcending geographical boundaries and fostering a global network of shared experiences.

These online communities experience various forms of toxicity, ranging from hate speech to cyberbullying. Volunteer moderators are frontline guardians in these communities, playing a pivotal role in maintaining their safety and well-being by fighting hate and harassment. Community moderators dedicate countless hours to enforcing community guidelines, curbing the spread of harmful content, and sanctioning offenders. On Reddit, for example, moderators provide, on average, $3.4 million worth of unpaid labor each year [29].

Volunteer moderators face various challenges in managing their community, including being personally targeted by harassers on the internet [5, 38], emotional burnout [15, 47], and lack of support from the platform [16]. Yet, there is little research examining how volunteer moderators seek out support in navigating through these challenges. While a few research studies looked at how moderators within a team collaborate to manage their community, revealing that they share frustrations and seek advice and affirmation on their actions from their peers [11, 18, 46], in this work, we examine what we term *moderator support* communities: online communities where moderators come together to discuss issues they encounter while moderating their communities and to request and receive support and advice from one another. These communities allow moderators to address moderation issues, with engagement from a diverse moderator community with different backgrounds and expertise.

In this paper, we complement prior works by examining mutual support among Reddit moderators in these moderator support communities: r/ModSupport and r/modhelp. We specifically focused on mutual support around fighting hate, harassment, and abuse, as it directly affects the moderator's

ability to keep the members safe and maintain a secure and respectful online environment. Both in these forums, Reddit moderators share topics or questions related to moderation to seek insights and advice from both their peers and Reddit administrators. We systematically analyzed moderators' discussions in these subreddits to understand:

- **RQ1**: How do moderators use "moderator support" communities for support in managing community safety? On what kinds of online hate and harassment topics do moderators ask for help or advice?
- **RQ2**: What types of advice are given in these communities by other moderators?
- **RQ3**: What role does this type of moderator-to-moderator support play in moderators' ability to protect the safety of their communities?

To answer these questions, we conducted a qualitative content analysis of 2,740 comments in over 115 threads about hate, harassment, and abuse, drawn from r/ModSupport and r/modhelp. Our analysis contributes to the Human-Computer Interaction (HCI) and Usable Security and Privacy research community in several ways:

- We are the first to our knowledge to provide a detailed characterization of different types of support exchanged among moderators in 'moderator support' communities to fight hate and harassment in online communities.
- We offer implications for design to facilitate peer and platform support for online community moderators in managing their own and community's safety and protecting them from online harm..

2 Related Work

Volunteer moderation in online communities

Moderation in online communities can be defined as "the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse" [20]. Volunteer moderators are the main drivers for moderation in many social media platforms, like Reddit, Facebook, Twitch, etc. These moderators wear many hats. They act as the custodians of community rules, explaining them to newcomers and reminding established members by enforcing the norms and guidelines [19]. They're also detectives, identifying violations and rule breakers and taking action by removing contents and punishing the violator [10]. But the moderator's job isn't just about enforcing order; they're also cheerleaders, fostering a sense of belonging and encouraging participation [46, 51].

Researchers have explored volunteer moderation across different platforms, offering varied insights into their roles and experiences [5, 10, 16, 24, 46]. Several looked at the digital labor of volunteer moderators and found that they spent a significant amount of time ensuring their community's growth

and safety [29, 33]. In an interview with volunteer moderators, Dosono et al. found that moderators spend, on average, 2-3 hours daily managing and moderating their respective communities [15]. In addition to manual labor, they also shed light on the emotional labor moderators endure dealing with hate, harassment, and negativity in their subreddits. Steiger et al. shared the same sentiment, emphasizing the psychological impact of moderation in establishing and preserving personal boundaries to avoid burnout and navigating complex interpersonal conflicts within the community [47]. Schöpke et al. pointed out that in addition to disgruntled community members, psychological distress also stems from struggles with other moderators in the team [44].

Prior research with moderators also demonstrates the challenges moderators face in balancing free speech and community safety [25, 34], providing transparency of moderation decisions [7, 22], effectively communicating and collaborating with the moderation team [10], and with the moderation tools lacking the nuance required in considering contextual factors and corner cases [19, 23, 27]. Some others shed light on the challenges associated with moderation strategies based on interaction mediums, such as in voice-only communities in Discord [24] and live-streaming communities like Twitch [50]. Despite the vast majority of work looking at the challenges moderators face, little is known about the support moderators employ to navigate through these challenges. We extend the current literature by particularly looking at mutual support among moderators in managing community safety.

Peer communication and support in moderation

Multiple studies have investigated how social media platforms are used for general support-seeking, such as in navigating unemployment [17], job loss [8], dealing with specific physical health conditions [4, 35, 42], mental health [14, 37], or the death of a family member [6]. However, these studies also indicated that these spaces could lead to negative experiences, i.e., aggressive content, stalking, exploitation of shared information, etc.. These studies emphasized the support that online moderators provide to minimize such activities within their communities [2, 26, 41].

Some studies explored how moderators within a team work together to maintain supportive and safe online spaces. Chi et al. examined the communication and collaboration methods employed by volunteer moderators on Twitch, emphasizing the importance of both informal and formal communication to facilitate teamwork among moderators and streamers [11]. Seering et al. investigated how moderator teams interact in community development, observing that team members often discuss specific incidents to solicit advice or opinions on the most appropriate course of action, as well as to inform other moderators about actions taken [45]. In an ethnographic study involving Facebook moderators, Gibson et al. discovered that some moderators view their team as a source of validation and

confidence in their decisions, which helps alleviate the anxiety associated with volunteer moderation [18]. Moderators also express their frustrations within their moderation teams as a means of seeking social support [15, 18]. Beyond internal collaboration, moderators from various subreddits on Reddit joined forces in a large-scale collective action by temporarily shutting down their subreddits, demanding improved support from platform administrators [32, 39].

While most previous works have focused on moderators' reliance on their team for guidance and assistance, this paper explores mutual support among moderators beyond their immediate moderation team. Our research offers a comprehensive overview of how moderators leverage the expertise and experience of the broader moderator community to address challenges related to combating hate and harassment within their community.

2.1 Moderation in Reddit

Moderation on Reddit operates through a combination of automated tools, subreddit moderators, and platform administrators. Each subreddit is overseen by volunteer moderators who enforce community guidelines by removing inappropriate content, issuing warnings, or banning users who violate rules. Reddit also employs a team of paid administrators who manage site-wide policies and legal matters and can issue site-wide bans when necessary. Communication between users and moderators is facilitated through Modmail, a shared inbox where users can report violations and moderators can address community concerns. Automated tools such as "automoderator" assist moderators in identifying, filtering and removing abusive content. The "modqueue" serves as a central hub within each subreddit, listing all content pieces that needs moderator review, including user reports, filtered posts, and comments.

3 Methods

In this section, we present details of the data collection, filtering and the data analysis procedure.

3.1 Data collection & sample generation

In our study, we focused on two subreddits, r/ModSupport and r/modhelp. These platforms are specifically created for moderators to engage in discussions covering a wide range of topics, such as moderation issues, tools, and instances of online abuse within the community, etc. Moderators utilize these forums to seek assistance and guidance from both administrators and fellow moderators. These two communities have substantial user bases: r/modhelp, established in 2009, is the largest moderator community on Reddit with 121k members, while r/ModSupport, established in 2015, has 72.8k members as of February 2023. Both subreddits exhibit high activity

with over ten daily posts (original submission made by a user in a subreddit), providing a rich dataset for our research. We downloaded all available posts from these subreddits from inception to December 2022 from pushshift.io [40], a platform that is used to maintain an up-to-date public archive for Reddit. We omitted posts where the that were empty, deleted by the poster, or removed by moderators. The resulting corpus contains 41,256 posts.

We focused on posts where moderators discussed hate, harassment, and abuse-related attacks towards their community or themselves and/or were asking questions/suggestions about those and sharing challenges in keeping their community safe against those attacks. We used a broad definition of hate and harassment taken from Pew Research [3] and Thomas et al. [49]: *"Hate, harassment, and abuse occur when an aggressor (either an individual or group) specifically targets another person (including moderators) or group to inflict harm: emotional, financial, or physical. In its milder forms, it creates a layer of negativity that people must shift through as they navigate their daily routines online. At its most severe, it can compromise users' privacy, force them to choose when and where to participate online, or even pose a threat to their physical safety, e.g., doxing and swatting."* Through an iterative process, we developed a set of 35 keywords and key phrases drawn from Thomas et al.'s taxonomy [49] to which we added the set of reasons that Reddit's report form offers users for describing content that breaks site rules to identify posts relevant to hate, harassment, and online abuse, provided in Appendix A.1. We searched the posts containing these keywords, which left us with 3,321 posts in our final dataset. More details about the selection of subreddits and the process of generating keywords can be found on [48].

Final Sample: Two researchers randomly sampled a post from the dataset. They manually reviewed and discussed the post. If the post was unrelated to online hate, harassment, and community safety, it was considered a false positive and replaced with a new, randomly sampled post. Otherwise, they downloaded and coded the entire thread (the entire discussion that unfolds from the post, including the post, all the subsequent comments, and replies) associated with the post. This process of random sampling and coding continued until we reached saturation, following the guidelines in prior research [43] In total, we coded and reached saturation with 115 relevant threads (2740 comments) sampled from our final dataset. The sample is also used in another paper of our authorship [48] to systematize adversarial attacks on Reddit that are happening by exploiting platform features and identifying challenges moderators encounter for such exploitation. In this paper, we scoped our analysis to understand mutual support among moderators in mod-support communities to fight hate and harassment in the community they moderate, which was not investigated and reported in [48].

3.2 Data analysis

The goal of the data analysis was to evaluate each thread in the sample for the type of requests in the original post and the support exhibited in the associated comments. We used an open coding process to identify the types of support sought by the moderators in the posts. To develop the codebook for the types of support exhibited, we drew from the offers and provisions of support, as defined by Cutrona and Suhr's Social Support Behavioral Code [13], and adjusted our codebook to include only support codes and subcodes present in our dataset. Additionally, we added new codes and subcodes that emerged from our analysis that were not present in Cutron and Suhr's Support Code. For instance, we have added a 'clarification' subcode under the information support code and the 'unsupport' code in our codebook. This process consisted of having three researchers go through 50 threads in multiple rounds to reveal initial codes. The research team met multiple times in this process to discuss the codes, clarify definitions, resolve disagreements, and establish an initial codebook. Two researchers then coded sets of 20-25 threads at a time, meeting between sets to compare codes, resolve disagreements, and revise the codebook until no new code emerged. Both coders coded and discussed the same set of threads and agreed on the codes, so we do not report inter-coder agreement. Finally, the codes were grouped into categories in order to characterize the kind of support requested and received presented in section 4. The research team held regular meetings to review and discuss the analysis results and the categories generated from the analysis.

3.3 Ethical considerations

Our institution's Institutional Review Board determined that this study was out of scope for their oversight. Nevertheless, this work has significant ethical implications for the moderators whose words we studied and the communities they protect. The data we analyzed are direct quotes from Reddit moderators, many of whom are from marginalized communities. Though this content is publicly available to anyone, our aggregating it as a dataset and highlighting aspects of it in this manuscript could induce unwanted or dangerous attention (including hate and harassment) towards moderators and their communities. We took several steps to mitigate these dangers. We chose not to release the aggregated dataset publicly. We redact any usernames, specific subreddits (other than `/r/ModSupport` and `/r/modhelp`), or specific communities (e.g., when discussing subreddits associated with a physical city). Additionally, to increase the difficulty of re-identifying specific posts, comments, posters and commenters for targeted harassment, we have paraphrased all quotations that appear in the paper (one researcher paraphrased and another reviewed each paraphrase for fidelity to the original meaning).

3.4 Limitations

In this research, we only focused on public-facing posts and comments on Reddit specific English-language subreddits, within the context of mutual support in managing community safety. As such, it is uncertain how applicable our findings are to other platforms such as Facebook, Twitch, Discord, etc., which may have different moderation structures or to other contexts that are not related to hate and harassment. Moreover, moderators on Reddit may also seek support in other ways, such as in private subreddits or channels, which are not covered in our work. The aim of this study is not to establish generalizability but rather to examine a specific phenomenon within a particular context. Due to the nature of our research, many variables remain unknown. We do not have access to any data regarding the demographics of the moderators we studied, leaving their gender, education level, occupation, age, and location undisclosed.

4 Results

This section presents the results of our analysis of the Reddit data from two moderator support communities, r/ModSupport and r/modhelp, in reference to our research questions. In the rest of the paper, we use the following terminology:

- 'Community' refers to communities/subreddits moderators moderates unless otherwise specified.
- 'Moderators' or 'posters' indicate moderators from various subreddits who posted to seek assistance in moderator support communities.
- 'Commenters' or 'Redditors' are individuals who responded to these posts.
- 'Admins' are Reddit administrators unless otherwise specified.

4.1 Purpose of posting

Moderators who sought support tended to engage with the moderator support community, reaching out to everyone for guidance. They discussed various issues, such as the hate or harassment they encountered, the difficulties in ensuring the safety of their community, or simply expressing their frustrations to be acknowledged. They openly complained about their problems, shared personal experiences, and recounted specific instances where they explicitly sought support and advice. We have found that the support requested by the moderators falls into three major categories: suggestion/advice (63 threads); clarification of platform, features, and rules (31 threads); and feature/tool requests directed at Reddit administration (13 threads). These categories are not mutually exclusive, and some posts fall into multiple categories. When asking for support for any of these categories, moderators often shared frustration with their situation and challenges

(30 threads). In fourteen threads, the poster did not explicitly ask any questions or request any specific tool/feature; instead, they simply expressed their frustration with admins, AEO (Reddit anti-evil operations team that identify and address violations of Reddit's policies on the platform), or Reddit (i.e., lack of response from admin, wrongful action by AEO, etc.) and the platform's lack of support. Table 2 in Appendix A.2 displays the types of support requested by moderators in our sample with examples. In the subsections below, we describe each type of support seeking in more depth.

4.1.1 Requesting suggestion/advice

In most posts within this category, moderators sought suggestions and guidance on combating instances of hate and harassment they were experiencing (50 threads) or anticipated in the community they moderate (5 threads). The most common issues moderators faced were attackers spreading hate and harassment via mass downvoting, false reporting, spamming, harassing posts/comments/PM, etc., and a lack of support from Reddit in fighting these attacks. Sometimes, these attacks are specifically targeted to moderators of the community. For instance, the moderator mentioned: *"One user is personally attacking one of our moderators. They have even created a username with her name and the C word. He targets any comment she makes on the internet and says he will continue to escalate until she deletes her Reddit account."*

Moderators' inquiries spanned from seeking advice on handling specific attacks to asking for feedback on the actions they had taken or intended to take to stop or prevent such attacks. For instance, one moderator mentioned: *"My subreddit's theme attracts nazis, racists, and transphobic. I have issued numerous bans and begun taking mod applications to have more help dispelling this type of behavior. Yet, I doubt that addressing this issue through bans will actually solve anything. Are there any mods who have effectively prevented rampant bigotry in their community? How is it done?"*

In a few instances (4 threads), moderators asked for guidance regarding how to help a community member at risk. These community members were in vulnerable situations, like being abused or threatening suicide, or receiving targeted attacks like doxing and revenge porn. Moderators used the resources they had at hand, like reporting to Reddit and sharing supporting resources with the user. However, in all cases, we observed moderators feeling responsible for community members beyond that and asked for advice from other moderators. For instance, one moderator said: *"A user in my sub seems to be in an abusive and threatening situation. I understand that they need help, and I would use the Endangerment or suicide/self-harm form to report it, except that would not fit this circumstance. Also, I am outside of the US so I can't report this to my country's authorities. Do I need to contact (Reddit) support so that they may disclose the user's location to the police? How would I do that?"* Finally, in a few other

threads (4 threads), moderators asked for advice on addressing wrongful actions taken by the AEO, best practices for adding new mods and managing their own safety.

4.1.2 Requesting clarification

Within the support-seeking post, 31 were posted where moderators posing queries or seeking clarification on various matters. Seventeen threads indicated moderator confusion regarding platform functionalities and features including how a specific feature works, appropriate ways to use a feature, the differences between features, when to use which feature, and how the platform makes decisions. Examples include inquiries like, can mod log be edited? what are the differences between different types of bans? which reporting category to use to report a particular instance? how AEO works?, etc. For instance, one moderator sought clarification from the admin, stating, *"Admins should disclose to us why they delete posts and comments. we should be notified with a simple message indicating violence, threat, dox, just something to guide us so we can better moderate according to their standards."*

In some other posts (12 threads), moderators sought clarification around Reddit policy and rules around specific matters(i.e., what constitutes brigading? What is the policy around doxing? etc.). For instance, one moderator asked: *"Recently, we have received multiple requests from a certain company to delete confidential data (email threads, sales reports, email addresses). I don't want to remove these because the posts are well-written and produce healthy conversation; they just attach a photo of information that this company does not want to be public. Am I legally required or under any obligation according to Reddit's terms to delete these posts?"*

Finally, in two threads, moderators asked for clarification on the modmails they received from Reddit to understand why they had received such warnings.

4.1.3 Tool/feature request

In several threads (13 threads), moderators specially requested tool or feature support from the admins. In most of the threads in this category (7 threads), moderators asked for features/tools to prevent attacks against the community. For instance, once moderators requested a tool to automatically flag and remove the accounts that evade bans and automatically remove posts/comments from ban evaders and block them from doing further harassment. Another moderator wanted the feature to only allow subscribers (who were subscribers on or before a certain date) to comment on a post to reduce the risk of brigading. In four of the threads, moderators requested features to help them detect and report offenders (i.e., the ability to see edited comments, the ability to add explanations while reporting, etc.). In a few threads (2 threads), moderators specifically asked for features to reduce moderators' harassment, such as the ability to seamlessly switch between

their dedicated moderation account and regular user profile, concealing moderators' identities when interacting with rule violators, and limiting the number of modmails someone can send to the mods in a certain period. One moderator stated:

"We were just inundated with harassing messages from a single user sending 30 messages in a minute, which I did not realize was possible. I cannot comprehend any circumstance in which this spam messaging would be wanted or acceptable, so why not take away someone's ability to do this?"

In summary, moderators use r/ModSupport and r/modhelp as a space for expressing and discussing various issues in the context of their experience fighting hate and harassment in their subreddits. These discussions encompassed topics such as attack prevention techniques, the safety of community members, moderation tools and platform policy, moderators-admin communication, platform moderation, and automated tools used by platforms. In our sample, most moderators posted using the account they were using to moderate. Some moderators who were facing targeted harassment used throwaway accounts to ask for advice to prevent stalking. Using the mod account is not surprising as these subreddits are specifically created for moderators and overseen by administrators, likely fostering a level of trust among posters despite occasional rude comments. Moreover, moderators often needed to discuss specific challenges encountered while moderating particular types of subreddits as exemplified by statements like *"I manage a subreddit focused on mental health, and there's a user actively encouraging our suicidal users to commit suicide."* Interestingly, some moderators believed they faced specific difficulties due to moderating niche communities: *"truly feels like Reddit admins ignore the NSFW community. Our issues are falling on deaf ears."*

4.2 Support received

We observed a wide range of types of support offered in response to moderators' requests and the challenges.

Most of the comments were positive, providing information, clarification, and validation to moderators. They often engaged in meaningful discussions, asked follow-up questions, and expressed gratitude. We observed only a few instances where commenters exhibited unsupportive or negative behavior. However, the fact that we rarely observed negative behavior might be explained by it having been effectively moderated, since it would be reasonable to expect these moderator-focused communities, one of which is moderated directly by Reddit administrators, to be promptly moderated.

We identified four categories of support among supportive comments: informational, validation, emotional, and instrumental. We describe negative behavior under a separate 'unsupport' category. Table 3 in Appendix A.3 depicts the types of support exhibited in moderator support communities in our sample with definitions and examples. We characterize these types of support in detail in the following subsections.

4.2.1 Information support

Cutrona and Suhr defined information support as providing information about the problem or how to deal with that [13]. In mod support communities, Redditors shared insights, ideas, and suggestions with fellow Redditors, aiming to assist them in better understanding their situations and making more informed decisions. Redditors provided information support by providing strategic advice, clarification and explanations, assessment of the situation, and referral to other people and resources. 108 threads out of the 115 threads received some form of information support from fellow Redditors.

Strategic advice: One of the main ways information support is demonstrated is through offering suggestions and advice [13]. In our dataset, Redditors provided strategic advice to handle ongoing attacks, how to prevent future attacks, and how to protect moderators and their communities from hate and harassment. We observed such a form of support in 61 threads. Sometimes, Redditors provide direct advice applicable to the posters' situation. Other times, they shared experiences of handling a similar situation and their personal moderation practices instead of giving direct suggestions to the posters. For instance, one Redditor shared their moderation practice in response to a moderator experiencing Brigading: *"Crossposting is the root of brigading and that is the main issue. We impose bans on xposters and lock and remove their posts. This process is written in our sidebar. We know it is controversial, but has functioned effectively in our community for years. The number of users from other subs who crosspost or abuse our sub has decreased significantly."*

These suggestions, however, are not always unanimous, triggering back-and-forth discussions. For example, a query about how to handle a troll induced the following discussion:

"C1: Document all of their actions in the next few months. Take screenshots of everything they do. Post all this evidence when they come back to disparage the mods, then ban them. C2: Too much effort. Take a firm stance. If they post another controversial comment against the mods, give them a temporary 30-day ban. Explain in the ban comment space that their trolling is not welcome in your sub and that the next time it happens, however minor, they will be permanently banned. Be banned or behave—those are their options."

Redditors frequently suggested configuring auto mods (i.e., restricting new accounts, screening new users, filtering posts and comments based on keywords, etc.) to defend against harassment and attacks. Other suggestions include strictening the community rules, actioning offenders (deleting posts/comments and ban), adjusting features (i.e., making the subreddit private, only allowing pre-made fliers, etc.), using existing third-party bots (i.e., safest bot, totesmessengerbot, etc.) or developing custom bots to detect and defend against abuse. Furthermore, there were a few instances where commenters recommended some best practices to reduce harassment and ensure posters and their community's safety, such

as using separate accounts for modding and general Reddit use, using a throwaway account when asking questions about something delicate about the community or community members, etc. For example, in response to a moderator sharing his struggle with a stalker, one Redditor suggested: *"Use a separate reddit account for posting outside of your subreddit. Only use your existing Reddit account for moderating in your subreddit(s). Other mods have spoken of doing this even if they haven't been harassed to avoid potential stalkers. It is impossible for stalkers to track you if you do not have posts outside of the sub(s) that you moderate."* Another Redditor suggested: *"Delete comments indicating suicidal intentions after sending the self-harm report to prevent any potential bad actors from contacting the user privately and exacerbating the situation."* to a moderator dealing with a suicidal user.

Situation assessment: According to Cutrona and Suhr, information support can also be offered by helping support requesters reassess or redefine their situation [13]. On 46 threads, we observed Redditors providing support by assessing someone's situation from their experience with Reddit and moderating communities. It includes analyzing why and how an attack or harassment may have happened, why admins or AEO may have taken a particular action, why someone's approach to handling an attack is working or not working, etc. For instance, one moderator posted about receiving *"anti-semitic language, as well as violent threats and homophobic and transphobic language"* attached with the reports in his subreddit, and one commenter assessed the attack by saying: *"most likely these are deliberate trolls attempting to get a reaction from someone. It's possible that all of the messages are from one user trying to appear a bigger threat than they are. Unless you have given them any personal info it is highly unlikely that they will follow through on any of their threats."*

In another instance, one moderator was concerned about the trolling attack with new accounts and shared their approach to handling trolling, and a commenter assessed why their approach failed to stop the trolls. He said: *"It's completely reasonable to ask an account to be a week old before posting. It gives genuine users a chance to become more familiar with your community before actually posting. You are being deceived in two ways: actual new users in your subreddit must both be new to Reddit or use a pseudonym, and they must have a topic so urgent that they cannot wait your probationary period to post it."* Similar to the strategic advice and clarification category, assessments were not always uniform and triggered back-and-forth discussions.

Referral: Cutrona and Suhr describe referral as directing an individual to other sources of help [13]. On 78 threads, we have observed at least one Redditor referring poster to other sources who can help with their specific queries or issues.

Referral to admin: For 71 out of the 115 threads, at least one commenter referred the person facing specific challenges to admin to solve their problem. Moderators were referred for report abuse, ban evasion, organized harassment/brigading,

targeted and persistent harassment, and stalking. The high volume of referrals to Reddit admin is not very surprising, as we observed that the moderators often were unable to prevent and defend against those attacks with the tools and power they had at their disposal. The commenters also echoed the same challenge when they referred the poster to the admin.

Commenters offered such support by providing the link to the report form and message link to the admin, providing ideas on the option to select from the report form to reflect OP's situation and the information to include in the report. For instance, in response to how to modmail admins about username abuse, one Redditor suggested what information to include in the modmail to report abuse of award feature: *"Be sure to report the harassing username of the award giver. In no way, mention the username of the awardee and also give them a link to the harasser's Reddit profile. Don't report it in case someone commits an error."*

Moderators were advised to begin by reporting encountered issues using the platform's report form. However, if the individual who made the report did not receive any response, experienced inaccurate actions taken by Reddit, found that their efforts to address the abuse were ineffective, or if the situation was urgent, such as ensuring the safety of a community member facing issues like suicide threats, being in an abusive relationship, or encountering pedophilic activity, moderators were advised to directly message the admin via modmail.

We have observed a genuine effort from volunteer moderators to capture the admins' attention, even if that requires more time and energy from them, and that was reflected when they were providing advice. For instance, some Redditors suggested reporting every single example of report abuse and ban evasion even though it is time-consuming to report large scale attacks manually. Some also suggested not deleting the reports from the mod queue to protect the evidence, although it clogs up the mod queue and makes moderating problematic.

Referral to others: In the extreme cases of stalking and harassment like doxing and physical threats, the posters were referred to the legal authorities, e.g., the police and FBI. Some commenters also explained the law and how to approach the authorities: *"In Australia, it is a federal offense to use an online service to harass, threaten, or be offensive with a punishment of up to 3 years in prison. If you can locate his state, call that state's police (even though it's a federal crime, it is too minor to be enforced by the Australian federal police.) and initiate a report with all screenshots."*

Moderators were also referred to other subreddits and people who could help them with their issues. For instance, legal help subreddits for how to take steps against harassment, suicide watch-related subreddits for resources on how to help a user, automod-related subreddits to help set automod and code custom bots, mod reserves to get additional moderators in an emergency, etc. For example, in response to a moderator struggling with mass downvoting, one commenter suggested: *"I know of a subreddit called /r/x that used automod settings to*

prevent downvote brigading, maybe they could weigh in? You could reach out to their mods?" A few times, Redditors suggested seeking support from their community by exposing the abuse to the community. For instance, one Mod complained about another sub plagiarizing their post, and someone suggested: *" I'd recommend alerting your sub members to the situations and have them report as soon as it happens."*

Clarification: In addition to Cotrona and Suhr's information support behavioral codes advice, situation assessment, and referral, we have observed Redditors providing information that clarifies someone's confusion or misconception about the Reddit platform, features, and policy and was observed in 61 threads on our dataset. Redditors help moderators by explaining how the Reddit platform and different features work, what the rules are surrounding online attacks and harassment and how to report those to the Platform administrators. For instance, to clarify a moderator's confusion about *"Regarding doxing, where is the boundary between an influential individual's Twitter/social media and the socials of a local small business owner?"*, one Redditor explained: *"A published tweet isn't doxing. An example would be to say 'the individual that did x lives at 123 x avenue and their name is Jane Doe, here is a photo of them and phone number!"'*

However, these clarifications often come from the commenters' experiences and understanding of Reddit and may not match with reality. We have observed conflicting conversations among Redditors on how something works. For instance, two Redditors were in conflict about the difference between admin removal and moderators' removal of posts:

"C1: When it's deleted by the admin... it's deleted completely from the site. When it's deleted by a moderator, individuals can still view it if they use a particular link.
C2: To the best of knowledge, a deletion done by a moderator is the same as a deletion done by the admin.
C1: That is dependent on how they delete it. Admin can remove content off the entire website. Moderators can only obscure it from general users.
C2: I've looked into this and I'm afraid you're wrong."

4.2.2 Validation support

Another form of support we observed is validation support (on 50 threads), where Redditors validate posters' experience and/or needs. Cutrona and Suhr describe validation as a way of providing esteem support, i.e., communicating confidence by validating the recipient's perspective regarding a situation [13]. We included validation as a separate support code due to its prevalence in our dataset and established two ways validation is offered: confirming recipients' experience (i.e., confirming frustrating situations with AEO and Admin, confirming facing similar attacks and abuse, etc.) and endorsing suggestion/request (i.e., showing agreement that the requested tool or clarification is needed, endorsing someone's suggestion to prevent an attack, etc.). The validation support com-

ment generally started with sentences like *'can second this'*, *'I have the same experience'*, *'I think that would be super helpful as well'*, etc. For instance, in response to someone sharing their concern about the increase in the frequency of abuse during Pride month, someone replied: *"This conversation has been very valuable. We've already experienced a spike in reporting for "misinformation" or "harassment" on any post with gay or trans content."*

In a few instances, we have observed Redditors validating platform updates. For instance, one moderator mentioned: *"I love the new blocking feature. The improvements they have made to blocking have greatly improved my quality of life. I find myself being stalked less frequently around Reddit."*

4.2.3 Emotional support

Curtona and Suhr defined emotional support as the provision of love, care and empathy [13]. Redditors provided emotional support by showing their fellow Redditors appreciation, care, understanding and encouragement on 27 threads. Redditors appreciated the moderators for their work and for managing particular subreddits that help users in need. For instance, one moderator shared his experience of harassment running a local subreddit during a mass shooting in that area. To appreciate the moderator, one Redditor responded: *"Y'all performed wonderfully in the aftershock of that catastrophe. Thank you for your work."* The moderator expressed gratitude by saying: *"Thank you for the acknowledgment and your gracious words. Your reply got a bit buried but even if it's just to let us know we are on the right track, I really value the response. "*

Redditors showed sympathy and care to fellow Redditors, especially when someone harassed or stalked online. For example, in response to a moderator sharing harassment experiences, one Redditor said: *"I am so sorry you were put through all of that pain. You're only trying to assist folks. You're a really great person, I don't believe I would continue moderating a sub with users that harassed me or others in that way."*

Redditors also provided emotional support by showing understanding of fellow Redditor's emotional condition. For instance, one moderator shares his/her experience helping a suicidal person while having a suicidal tragedy in the family. One Redditor empathized by sharing his own experience: *"I feel you. I nearly went through that with a loved one. They only began speaking to me again after their failed attempt. Before that, I had attempted to help her numerous times through panic attack. But we made no progress. Things only changed after she opened up again, it was horrific."*

4.2.4 Instrumental support

Instrumental support is comparable to providing 'tangible assistance' by offering goods and services [13]. Beyond just offering advice and knowledge (information support), we observed Redditors in our dataset providing tangible resources

or offering to provide specific services to help with the recipient's situation (38 threads).

The tangible resources include automoderator codes, custom bots, and materials that could potentially solve OP's problem. For instance, in response to a moderator asking for advice to clean up Bigotry in their subreddit, one Redditor responded: *"If it would help, there is an extensive filter in /r/<redacted> to stop *a ton* of slurs and bigoted terms. I am happy to share the Automod code to you."* However, we have observed instances where Redditors could not share resources that they think could be helpful to the poster. For instance, in response to a moderator sharing their struggle fighting with a well-known spam bot, one Redditor wanted to share a document that described the individual behind it and their strategy, various websites and domains they owned, and other details. However, they later realized the document was from a private Reddit community and said: *"I'm unable to even archive the page. I can attempt to contact one of their moderators and see if they'll allow me to share information regarding the post."* In a few cases, Redditors provided instrumental support by sharing their willingness to get directly involved in the issue posters face and trying to find a solution. These include offering to collect information for the poster and talk over DM to help build solutions. For instance, one Redditor wanted to set up an automated bot to help suicidal users with resources, and another Redditor showed support by saying: *"Let me know if you'd like me to do some more research for y'all :) Looks like you are already quite busy. "*

4.2.5 Unsupport

On 30 threads, at least one Redditor showed unsupportive behavior by questioning, demeaning, blaming, or bullying the poster or other commenters. Although several such comments were merely intended to insult or harass or were irrelevant to the thread's discussion, we noticed that certain unsupportive remarks served as a form of community self-moderation. In several instances, Redditors helped uphold community standards by criticizing posts or comments that violate guidelines, such as calling out someone for engaging in harassment within the thread. We also observed commenters holding posters accountable for their actions by citing their own experiences with the poster's subreddit or particular incidents moderated by the poster. For example, one moderator discussed conflicts with some users regarding a ban and expressed worry that it might escalate into a potential brigading attack on their subreddit. In response to that, one commenter responded with: *"Don't act like a victim, you messed up.".* Another commenter said in the same thread:*"Great. /r/x is a very toxic subreddit that bash men constantly. The sub's moderators are a complete disaster. I'm pretty certain u/x had an awful childhood experience that caused them to behave poorly towards men."* These initial comments sparked a cascade of criticism, where numerous commenters joined in to express their disapproval

of the moderator seeking support and adding negative remarks about them. Such a response from the mod-support communities could lead to more stress, exclusion, and burnout.

In several instances, we have observed the moderators of r/ModSupport and r/modhelp removing rude/harassing comments and/or banning the unsupportive Redditors. For example, one commenter was being rude and demeaning to the poster. Though moderators did not remove the comment, the commenter was temporarily banned with the statement: *"Your harsh words are absolutely unnecessary here. One mod is distressed and uncertain how to handle the harassment that they are enduring. I'll give you a three day ban from r/ModSupport. I hope this time allows you to think through some things."*

Summary: moderators primarily found support from their fellow moderators. Reddit administrators also engaged with 49 out of 115 threads by offering guidance and clarification. Nevertheless, in the majority of cases, they referred poster to the Reddit report form or suggested contacting them via modmail regarding the abuse rather than directly providing information or assistance. In a few instances, admins elucidated the platform's rules, explained how Reddit operates or delved deeper into the reasons behind the poster's specific problem. For example, in response to a user's complaint about wrongful suspensions of community members, an admin replied: *" Thank you for this post, and sorry you're frustrated. We have multiple teams that employ mixed methods of human review and automated tools to prevent offensive content from reaching Reddit users. However, both will commit the occasional error, just like mods do. For that reason we have appeals. I investigated the 3 suspensions you spoke of and in 2 of them (including the suspension of your fellow mod), were revoked through appeal. In the third case the suspension timed out on its own. "* Overall, admins displayed empathy towards the moderators, and in 29 of the threads where admins provided a response, at least one moderator expressed gratitude towards them. Conversely, in 21 instances, at least one commenter engaged in arguments and exhibited frustration with admin responses.

5 Discussion

5.1 Stress, coping and community support

Prior research has associated volunteer moderation with stress, trauma, and burnout [15,44,50]. In our analysis, we found that moderators emotional distress stemmed from personal harassment, secondary trauma resulting from combating harassment, and a lack of prompt assistance from platform administrators when needed". Dosono et al. found that moderators cope with emotional stress by "Building solidarity from shared struggles,"sharing frustrations with team members, or connecting with community members facing similar challenges [15].

In general, online groups can help individuals cope with a wide range of stressors by providing access to a larger and

more diverse community of support, compensating for the lack of support available in immediate social groups [12]. In this work, we've observed moderators utilizing the moderator support communities to cope with stress using two strategies following Lazarus and Folkman's transactional model of stress and coping [28]. Some posters managed their stress by seeking information and solutions to their problems that cause stress (problem-focused coping), others by expressing negative emotions such as sharing frustration and anger with their peers to manage their emotions (emotion-focused coping), and some by combining both approaches. Receiving instrumental and informational support in moderator support communities may assist moderators with problem-focused coping. Receiving validation, empathy, and care may help moderators with emotional-based coping to feel less isolated, especially when stress comes from immediate team members. In this work, we have observed commenters providing emotional and validation support even when seeking such support was not the primary intent of the post, helping moderators cope with stress and navigate the emotional labor they experience to sustain their community.

On the other hand, we also observed moderators receiving negative and unsupportive responses when seeking help. Moreover, moderators often referred to administrators for solutions, and their lack of response was a source of stress: "We've reached out to the admins but received no response yet. This situation is really stressing me out, which is the last thing I need during finals.". Not receiving immediate or effective feedback or encountering negative interactions within mod support groups may negatively affect coping with stress. Future work should investigate both the positive and negative influence of peer (un)support on moderators' stress management and its impact on community safety.

5.2 Empowerment through mutual support

Ammari and Schonebeck introduced the concept of networked empowerment that highlights how social media facilitates the process of empowerment through access to people going through a similar situation [1]. In the context of our study, network empowerment describes how moderators use mod support groups to find and learn from other moderators going through similar experiences, share resources to support moderation challenges, and empower each other through mutual support of insights, validation, and solidarity.

Networked empowerment is built on Zimmerman's model of psychological empowerment, which includes three components: the interactional component, the interpersonal component, and the behavioral component [52]. Moderators' discussions examined in our results can be mapped onto these components, allowing us to understand the roles that various types of support play in supporting moderators' empowerment and thus their work to protect and strengthen their communities.

The interactional component describes people's awareness and ability to act toward goals. In the moderator's support group, this component involved moderators receiving informational support and instrumental support from other moderators. Through this, moderators gain insights into various strategies for addressing issues, learn from other's experiences, discover resources for managing community safety, grasp a deeper understanding of platform features and policies, and broaden their perspectives by considering alternative viewpoints offered by experienced moderators. It could be particularly helpful in situations with solitary moderation or inexperienced moderator teams.

The intrapersonal component describes "how people think about themselves," which includes someone's perception of their ability to solve the problem at hand and perceived competence in taking the actions necessary to do that. Moderators pose questions to seek guidance from their peers and discuss their approaches to problem-solving. In return, moderators received informational, instrumental, and validation support that may help moderators increase their competence in dealing with the problem at hand. Prior research indicates that moderators feel more confident in taking moderation actions when they receive advice and affirmation from fellow team members [18]. Additionally, emotional and validation support helps moderators to think positively and cultivate a positive mindset, considering that their feelings and experiences are acknowledged and accepted by others. Conversely, we note the high frequency of threads (71/115) in which moderators are encourage to refer problems to admin. This may suggest that mods often do not feel empowered to solve the problem at hand through their own actions, and this feeling and its propagation through referral-type support may weaken moderator empowerment by weakening the intrapersonal component.

The behavioral components describe someone taking action to directly influence outcomes. This is demonstrated by moderators joining moderator support groups, asking questions, moderators assisting each other by providing answers, guidance, and resources, and even volunteering to directly address others' issues through informational and instrumental support. For instance, when one moderator expressed a desire to combat bigotry, another moderator offered to provide Automod code they use in their subreddit to filter out content containing slurs and other bigoted terms. This directly influences posters' ability to eliminate content containing bigoted language from their subreddit.

In addition to supporting moderators' personal goals, we observed moderators using mod support communities as a space to advocate for changes that would benefit the whole moderator community, by highlighting platform-wide issues, requesting tools and admin support, and discussing what is needed to empower moderators. In response, the moderators received support from the mod support community, validating said cause and strengthening their voices for change. While Ammari's model of networked empowerment highlighted how networks can empower individuals by supporting

their personal goals, our findings introduce another dimension: empowerment by driving changes that could help the entire network.

5.3 Design implications

Here, we present implications for design to facilitate mutual support among moderators and reduce the challenges they encounter in managing their personal and community safety.

5.3.1 Peer support

Develop a formal repository for moderators to share common and contemporary advice, resources, tools, etc. In online communities, moderators often encounter similar challenges or have similar goals for their communities. For instance, moderators from multiple communities may deal with the same spam bot or have the same goal of identifying community members who crosspost posts to trolling subreddits, etc. Redditors actively create content and mechanisms to address ongoing issues and share those with others to assist with moderation, automation, or fighting contemporary hate and harassment attacks [30]. Currently, these resources are primarily disseminated through informal channels such as word of mouth or direct requests for support, potentially leaving valuable resources unnoticed by moderators who could benefit from them. In some cases, individuals may need to reach out directly to the creators of these resources, as they are not publicly accessible. We understand that not all content can be freely shared due to security concerns—for example, sharing automod code used to safeguard against specific attacks could inadvertently expose vulnerabilities to adversaries. However, there are still valuable resources, such as guidelines to contact external support, guidelines for how to deal with suicidal community members, insights into an ongoing sitewide attack, tutorials about new moderation tools, etc., that could benefit other moderators without posing significant risks to anyone.

Reddit features a 'Wiki' section within each subreddit, enabling moderators and approved contributors to generate and collaborate on content. However, its utilization varies among different subreddits. While some wikis focus on elaborating the subreddit's rules or guidelines, others serve as repositories for FAQs or resources pertaining to the subreddit's topic. Even if a subreddit's wiki holds valuable content, individuals who could benefit from it may be unaware of its existence. The wikis of r/ModSupport and r/modhelp subreddit offer resources related to common moderation issues. Nevertheless, we observed moderators sharing useful resources to deal with contemporary issues that are not documented in these wikis. Establishing official repositories for moderators to share resources could ensure that valuable knowledge is accessible to all the moderators. However, such a repository could potentially open up a new avenue for malicious actors to target communities. For example, an adversary might create a tutorial for a tool and embed harmful code snippets within it. If a moderator lacking technical expertise were to execute this code, they or their community could inadvertently fall victim to an attack. Furthermore, someone could create a guideline document containing harassing language. We've noticed that resources shared in discussion threads often receive endorsements from multiple moderators, either through comments or by upvoting recommendations made to the poster. This serves to validate the reliability of the resources. The proposed repository should also incorporate a mechanism for users to endorse or dispute specific resources, enabling moderators to conduct their own assessments on whether and how to utilize said resources.

Develop mechanisms to facilitate building formal and informal relationships. Prior research has underscored the significance of receiving support from one's social circle, including family and friends, particularly for individuals employed in emotionally demanding fields [9, 21]. This social support serves as a means for individuals to express themselves, fostering a sense of belonging and comprehension, ultimately enhancing their mental well-being. Earlier studies revealed that volunteer moderators cope with emotional stress by conferring with fellow moderators within their respective teams [15, 50]. Our findings expand the prior research, indicating that moderators often alleviate frustration by venting within larger moderator communities comprising individuals from different subreddits who may be facing similar challenges. O'Leary et al. suggested that mechanisms that connect peers based on shared characteristics, beliefs, and needs can notably enhance peer-support [36]. We recommend that platforms should deliberately incorporate features to facilitate informal communication among moderators. For instance, platforms could establish official chat channels exclusively for moderators, enabling them to discuss issues, exchange experiences, and forge relationships in a casual setting. Additionally, providing moderators with the option to customize labels, where they can share information about themselves, such as interests, experience, the type of communities they moderate, etc., may facilitate the formation of sub-channels and lead to communication. Furthermore, establishing official mentorship programs where experienced moderators can guide and support newer moderators can help build relationships. Such initiatives can be particularly beneficial for novice moderators without experienced moderators in their team.

It is important to acknowledge that such features could potentially be exploited by adversaries to inflict additional harassment on moderators. Therefore, it's imperative that these support systems are meticulously designed, taking into account potential avenues for exploitation and incorporating robust defense mechanisms. Developing such features entails navigating complex challenges, and future research should explore support systems that enable moderators to connect and bond with one another in an unexploitative and safe manner.

5.3.2 Tool support

Establish an effective admin-mod communication structure. We have observed that moderators are frequently directed to contact platform administrators for their challenges due to a lack of sufficient tools for moderators to address these issues independently. However, moderators have encountered confusion and received unclear advice regarding what information should be included in their communication with admins for varying issues. On Reddit, moderators have the option to reach out to administrators by submitting report forms or sending modmail to r/ModSupport, a specialized channel designated for moderators. Yet, we have observed moderators frequently getting frustrated with admins not responding to them in a timely manner and, worse, receiving no response at all, even for critical issues like abuse of minors, suicidal users, etc. Previous studies also highlighted the lack of support from platforms in addressing moderation issues [16, 31, 45]. In 2015 and 2023, Reddit moderators went as far as participating in a blackout to demand support from platform operators and for additional moderation tools [32, 39]. Platforms should consider establishing an effective communication structure between administrators and moderators. Additionally, platforms should offer guidance on the format of this communication, specifying what information moderators should include in their report forms or modmails for particular issues. Platform moderators can employ automation to identify the most common issues they receive from moderators and provide communication templates for those. Furthermore, there should be a mechanism for platforms to prioritize time-sensitive issues where moderators require immediate attention. For instance, transmission of pornographic material to underage Redditors may require more immediate action than an accidental removal of a post by AEO.

Develop measures to empower moderators against personally targeted attacks. Prior research indicated that exposure to harassment is one of the primary reasons behind volunteer moderators quitting moderation [44]. In our analysis, we observed moderators seeking advice or support to cope with targeted personal harassment encountered while moderating their communities. In response, commenters suggested various best practices, such as using separate accounts for moderation and regular Reddit activities, refraining from sharing personal information on the platform, etc. Many of the moderators offering advice had themselves been victims of targeted harassment, including stalking and doxing. Notably, we did not come across any explicit mention of official guidelines for moderators instructing them on how to better protect themselves against such targeted personal attacks. We conducted an informal review of Reddit's moderator resources and found no specific guidelines on moderators' safety besides standard account

security recommendations such as utilizing strong passwords and enabling two-factor authentication (2FA). We advocate for the provision of training and resources for moderators not only on moderation techniques, tools, and protecting community members but also on self-protection measures against targeted attacks. Furthermore, as articulated by moderators in some of the threads we analyzed, platforms should consider introducing features that empower moderators to safeguard themselves, such as the ability to moderate anonymously using pseudonyms or anonymized profiles and seamless transitions between moderator and regular user profiles.

Adequately explain moderation features in place where moderators employ them. Our research reveals that moderators often seek clarification regarding the functionality of platform tools, features, or policies. However, they periodically receive conflicting answers, as responses from fellow Redditors may be based on assumptions rather than official information. Some clarification requires input from admins, like how AEO works. However, deciding to what extent such information should be shared is complex as it could be used by bad actors to evade detection. Then again we have also observed moderators seeking clarification about moderation features they use despite explanations being available in the official moderator help center. One reason could be that explanations are not readily accessible in the places where moderators apply these tools. Additionally, there may be insufficient detail provided about certain features. For example, while the Reddit moderator help center outlines activities a banned user cannot perform, it does not provide information about other features they can still use in the subreddit where they are banned. We suggest that platforms offer explanations directly in the places where moderators are more likely to use that information, covering all the details moderators may require about the tool through thorough user research. However, like any interface design, the challenge lies in providing sufficient yet concise information without overwhelming moderators.

6 Conclusion

Overall, our study provides a detailed characterization of the different types of support requested and received in the 'mod support' communities in the context of fighting hate and harassment. Our findings highlighted how support exchanged among moderators in these communities empowers them in managing community safety. The results unveiled implications around designing peer and tool support for moderators to better equip them to protect their own and community safety.

Acknowledgments

References

[1] Tawfiq Ammari and Sarita Schoenebeck. Networked empowerment on facebook groups for parents of children with special needs. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 2805–2814, New York, NY, USA, 2015. Association for Computing Machinery.

[2] Nazanin Andalibi, Oliver L. Haimson, Munmun De Choudhury, and Andrea Forte. Social support, reciprocity, and anonymity in responses to sexual abuse disclosures on social media. *ACM Trans. Comput.-Hum. Interact.*, 25(5), oct 2018.

[3] Sara Atske. The State of Online Harassment — pewresearch.org. https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/. [Accessed 02-16-2024].

[4] Ashley A Berard and André P Smith. Post your journey: Instagram as a support community for people with fibromyalgia. *Qualitative Health Research*, 29(2):237–247, 2019.

[5] Iris Birman. Moderation in different communities on reddit – a qualitative analysis study. 2018.

[6] Jed R Brubaker and Gillian R Hayes. " we will never forget you [online]" an empirical investigation of post-mortem myspace comments. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pages 123–132, 2011.

[7] Jens Brunk, Jana Mattern, and Dennis M. Riehle. Effect of transparency and trust on acceptance of automatic online comment moderation systems. In *2019 IEEE 21st Conference on Business Informatics (CBI)*, volume 01, pages 429–435, 2019.

[8] Moira Burke and Robert Kraut. Using facebook after losing a job: Differential benefits of strong and weak ties. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1419–1430, 2013.

[9] Carolyn M. Burns, Jeff Morley, Richard Bradshaw, and José Domene. The emotional impact on and coping strategies employed by police teams investigating internet child exploitation. *Traumatology*, 14(2):20–31, 2008.

[10] Jie Cai and Donghee Yvette Wohn. After violation but before sanction: Understanding volunteer moderators' profiling processes toward violators in live streaming communities. 5(CSCW2), oct 2021.

[11] Jie Cai and Donghee Yvette Wohn. Coordination and collaboration: How do volunteer moderators work as a team in live streaming communities? In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2022.

[12] Tsai-Yuan Chung, Cheng-Ying Yang, and Ming-Chun Chen. Online social support perceived by facebook users and its effects on stress coping. 2014.

[13] Carolyn E Cutrona and Julie A Suhr. Controllability of stressful events and satisfaction with spouse support behaviors. *Communication research*, 19(2):154–174, 1992.

[14] Munmun De Choudhury and Sushovan De. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 71–80, 2014.

[15] Bryan Dosono and Bryan Semaan. Moderation practices as emotional labor in sustaining online communities: The case of aapi identity work on reddit. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13, New York, NY, USA, 2019. Association for Computing Machinery.

[16] Bryan Dosono and Bryan Semaan. Decolonizing tactics as collective resilience: Identity work of aapi communities on reddit. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1), may 2020.

[17] Radhika Garg, Yash Kapadia, and Subhasree Sengupta. Using the lenses of emotion and support to understand unemployment discourse on reddit. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–24, 2021.

[18] Anna D Gibson. What teams do: Exploring volunteer content moderation team labor on facebook. *Social Media+ Society*, 9(3):20563051231186109, 2023.

[19] Tarleton Gillespie. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. 01 2018.

[20] James Grimmelmann. The virtues of moderation. *Yale JL & Tech.*, 17:42, 2015.

[21] Trond Idås and Klas Backholm. Risk and resilience among journalists covering potentially traumatic events. *The assault on journalism*, page 235, 2017.

[22] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. "did you suspect the post would be removed?": Understanding user reactions to content removals on reddit. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), nov 2019.

[23] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Trans. Comput.-Hum. Interact.*, 26(5), jul 2019.

[24] Jialun Aaron Jiang, Charles Kiene, Skyler Middler, Jed R. Brubaker, and Casey Fiesler. Moderation challenges in voice-based online communities on discord. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), nov 2019.

[25] Jialun Aaron Jiang, Peipei Nie, Jed R. Brubaker, and Casey Fiesler. A trade-off-centered framework of content moderation. *ACM Trans. Comput.-Hum. Interact.*, 30(1), mar 2023.

[26] Sarah Kendal, Sue Kirk, Rebecca Elvey, Roger Catchpole, and Steven Pryjmachuk. How a moderated online discussion forum facilitates support for young people with eating disorders. *Health Expectations*, 20(1):98–111, 2017.

[27] Tina Kuo, Alicia Hernani, and Jens Grossklags. The unsung heroes of facebook groups moderation: A case study of moderation practices and tools. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW1), apr 2023.

[28] Richard S Lazarus and Susan Folkman. *Stress, appraisal, and coping*. Springer publishing company, 1984.

[29] Hanlin Li, Brent J. Hecht, and Stevie Chancellor. All that's happening behind the scenes: Putting the spotlight on volunteer moderator labor in reddit. In *International Conference on Web and Social Media*, 2022.

[30] Kiel Long, John Vines, Selina Sutton, Phillip Brooker, Tom Feltwell, Ben Kirman, Julie Barnett, and Shaun Lawson. "could you define that in bot terms"? requesting, creating and using bots on reddit. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 3488–3500, New York, NY, USA, 2017. Association for Computing Machinery.

[31] Adrienne Massanari. #gamergate and the fappening: How reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*, 19(3):329–346, 2017.

[32] J. Nathan Matias. Going dark: Social factors in collective action against platform operators in the reddit blackout. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 1138–1151, New York, NY, USA, 2016. Association for Computing Machinery.

[33] J. Nathan Matias. The civic labor of volunteer moderators online. *Social Media + Society*, 5(2), 2019.

[34] Aiden R. McGillicuddy, Jean-Grégoire Bernard, and Jocelyn Cranefield. Controlling bad behavior in online communities: An examination of moderation work. In *International Conference on Interaction Sciences*, 2020.

[35] Hyun Jung Oh, Carolyn Lauckner, Jan Boehmer, Ryan Fewins-Bliss, and Kang Li. Facebooking for health: An examination into the solicitation and effects of health-related social support on social networking sites. *Computers in human behavior*, 29(5):2072–2080, 2013.

[36] Kathleen O'Leary, Arpita Bhattacharya, Sean A. Munson, Jacob O. Wobbrock, and Wanda Pratt. Design opportunities for mental health peer support technologies. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, page 1470–1484, New York, NY, USA, 2017. Association for Computing Machinery.

[37] Umashanthi Pavalanathan and Munmun De Choudhury. Identity management and mental health discourse in social media. In *Proceedings of the 24th international conference on world wide web*, pages 315–321, 2015.

[38] Benjamin Plackett. Unpaid and abused: Moderators speak out against reddit. https://www.engadget.com/2018-08-31-reddit-moderators-speak-out.html. [Accessed 02-16-2024].

[39] Jon Porter. Major reddit communities will go dark to protest threat to third-party apps. https://bit.ly/4fbH8ca, 2023. [Accessed 02-12-2024].

[40] Pushshift. Unpaid and abused: Moderators speak out against reddit. https://pushshift.io/signupl. [Accessed 02-16-2024].

[41] Koustuv Saha, Sindhu Kiranmai Ernala, Sarmistha Dutta, Eva Sharma, and Munmun De Choudhury. Understanding moderation in online mental health communities. In *Social Computing and Social Media. Participation, User Experience, Consumer Experience, and Applications of Social Computing: 12th International Conference, SCSM 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part II 22*, pages 87–107. Springer, 2020.

[42] Shruti Sannon, Elizabeth L. Murnane, Natalya N. Bazarova, and Geri Gay. "i was really, really nervous posting it": Communicating about invisible chronic illnesses across social media platforms. In *Proceedings of*

the *2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13, New York, NY, USA, 2019. Association for Computing Machinery.

[43] Benjamin Saunders, Julius Sim, Tom Kingstone, Shula Baker, Jackie Waterfield, Bernadette Bartlam, Heather Burroughs, and Clare Jinks. Saturation in qualitative research: exploring its conceptualization and operationalization. *Quality & Quantity*, 52, 07 2018.

[44] Angela M Schöpke-Gonzalez, Shubham Atreja, Han Na Shin, Najmin Ahmed, and Libby Hemphill. Why do volunteer content moderators quit? burnout, conflict, and harmful behaviors. *New Media & Society*, page 14614448221138529, 2022.

[45] Joseph Seering. Reconsidering self-moderation: the role of research in supporting community-based models for online content moderation. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2), oct 2020.

[46] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. Moderator engagement and community development in the age of algorithms. *New Media and Society*, 21(7):1417–1443, 2019.

[47] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J. Riedl, and Matthew Lease. The psychological well-being of content moderators: The emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[48] Madiha Tabassum, Alana Mackey, and Ada Lerner. Investigating moderation challenges to combating hate and harassment: The case of mod-admin power dynamics and feature misuse on reddit. In *30th USENIX Security Symposium (USENIX Security 24)*. USENIX Association, August 2024.

[49] Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley, Deepak Kumar, Damon McCoy, Sarah Meiklejohn, Thomas Ristenpart, and Gianluca Stringhini. Sok: Hate, harassment, and the changing landscape of online abuse. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 247–267, 2021.

[50] Donghee Yvette Wohn. Volunteer moderators in twitch micro communities: How they get involved, the roles they play, and the emotional labor they experience. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13, New York, NY, USA, 2019. Association for Computing Machinery.

[51] Bingjie Yu, Joseph Seering, Katta Spiel, and Leon Watts. "taking care of a fruit tree": Nurturing as a layer of concern in online community moderation. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, page 1–9, New York, NY, USA, 2020. Association for Computing Machinery.

[52] Marc A Zimmerman. Psychological empowerment: Issues and illustrations. *American journal of community psychology*, 23:581–599, 1995.

A Appendix

A.1 Keyword list to filter post related to hate, harassment and online abuse

Base word	Word forms
Bully	Bully, Bullying, Bullied, Bullies
Troll	Troll, Trolls, Trolling, Trolled
Profane	Profanity, profane, profaned, profaning, profanities, profanely, profanatory, profanes
Offensive content	Offensive content, offensive contents, offensive post, offensive posts, offensive comment, offensive comments, offensive word, offensive words
Threat	Threat, Threats, Threatening, Threaten, Threatens
Violance	Violence, Violent, Violently
Incite	incite, inciting, incites, incited, incitement
Harassment	Harass, Harasses, Harassment, Harassed, Harassing
Dox	Dox, doxxed, doxing, doxes, doxx, doxxing, doxxes, doxed
Dogpile	Dogpile, Dogpiled, Dogpiles, Dogpiling
Raid	Raid, Raids, Raiding, raided
Brigade	Brigade, Brigaded, Brigades, Brigading
Mass downvote	Mass downvote, mass downvoting, mass downvoter, serial downvote, serial downvoting, serial downvoter, mass downvotes, serial downvotes
Abuse	Abuse, Abusive, Abusing, Abuser, Abused, Abuses
Impersonate	Impersonation, impersonate, impersonates, impersonated, impersonating
Stalk	Stalk, Stalks, Stalked, Stalker, Stalking
(Sexual| sexualization) & (Minor| minors)	(Sexual| sexualization| sexually| sexualize) & (Minor|minors)
Personal information	personal information, personal info, private information, private info, confidential information, confidential info
Self harm	self harm, self-harm
Suicide	suicide, suicidal
Racism	Racism, Racist
Bigot	bigotry, bigot, bigots, bigoted
Transphobe	transphobe, transphobes, transphobia, transphobic
Homophobe	homophobic, homophobia, homophobes, homophobe
Scam	scam, scammed, scamming, scams, scammer, scammers
AEO	AEO, anti evil operation, anti evil operations, anti-evil operation, anti-evil operations
Ban evasion	(ban|bans)&(evade|evades|evaded|evading)
Hate speech, Hateful, Hatred, Non-consensual intimate media, Revenge porn, Denial of Service, Explicit content, Vote manipulation	

Table 1: Keyword list to filter post related to hate, harassment and online abuse

A.2 Support requested in mod support communities

Types	Subtypes	Examples
Suggestion/advice (63 threads)	Combating hate and harassment attacks (55 threads)	*"Recently, we've had many users from other subreddits harassing our users over chat. As the only chat mod, I can't monitor the chat 24/7. How can I flag potential bad actors in real-time?"*
	Helping community members at risk (4 threads)	*"A user is expressing suicidal thoughts. I've informed Reddit, shared support resources, and offered to link him with a crisis counselor. I want to reach out to local authorities, but don't know who he is. Any advice would be helpful."*
	Managing personal safety as moderators (2 threads)	*"I have another account besides the one I regularly use. Should I use it as my mod account? Any suggestions? I do not want members to dig up my old content or stalk me."*
	Best practices to add moderators (1 threads)	*"We're searching for a new mod for our subreddit. What best practices do you follow when adding a mod?"*
	Addressing wrongful action taken by AEO (1 thread)	*"I noticed some comments containing "f***" were removed by AEO. But they were translations of video dialogue, not hate speech. Is it safe to approve these comments since AEO removed them?"*
Clarification (31 threads)	Platform functionalities and features (17 threads)	*Are there any differences between auto-mod shadowban and mod-ban other than the difference in the offender receiving notification?*
	Reddit policy and rules (12 threads)	*"Would it be considered as doxing if someone receives a DM saying, 'I've looked through your post history. It won't be difficult to locate you.'?"*
	Modmail warnings received from admins (2 threads)	*"The subreddit I moderate just received a wanrning about promoting hate without any details. Why are we getting warnings for rule violations that we're obviously not doing?"*
Tool/Feature support (13 threads)	To prevent attacks against community (7 threads)	*"Users in our subreddit are being targeted by followbots with offensive names. Please implement a system to prevent such abuse, perhaps a cooldown period for following users."*
	To detect & report offenders (4 threads)	*"I frequently encounter people posting harassing comments and then editing their text back to normal. Please show editing history for mods."*
	To reduce mod-targeted harassment (2 threads)	*" In my opinion, Reddit should hide all mod names when we interact with users who are receiving a ban."*

Table 2: Types of support requested in the moderator support communities to manage community safety

A.3 Support exhibited in mod support communities

Types	Definition	Examples
Information Support (108 threads)		
Strategic Advice (61 threads)	Provide recipient advice or ideas to stop ongoing attacks or prevent future attacks	*"Adjust your spam filter to the highest setting and secure your accounts as much as possible. If you're very concerned, you might ask the head mod to temporarily revoke mod permissions for everyone else, although this probably won't be necessary here."*
Clarification (61 threads)	Clarify recipient's confusion or misconception about the Reddit platform, features, and policy	*"Even if you ban someone, they can still view, vote, and report."*
Situation Assessment (46 threads)	Assess why recipient's is experiencing a particular situation	*"It seems this individual is based outside of the US, and some foreign ISPs are known to be quite lax regarding their customers' activities."*
Referral (78 threads)	Refer the recipient to some other sources of help	*"Talk to automod coders. They can help you deal with this. "*
Validation Support (50 threads)		
Confirm experience (41 threads)	Confirm recipient's frustration, concern or challenge is valid as they experience the same issues	*"I reported a harassing comment and got the same 'it doesn't meet the requirements' nonsense."*
Endorse suggestion/request (13 threads)	Show agreement with recipient's suggestion or request	*" Yes please!!! It would be incredibly helpful to have a way to explain why something is offensive or promotes hate."*
Emotional Support (27 threads)		
Appreciation (4 threads)	Show appreciation for recipient's work	*"Your subreddit is a lifesaver for nearly every moderator."*
Care/ Sympathy (11 threads)	Express sorrow for recipient's situation	*"I'm sorry that you had to experience this. Everyone deserves to be treated with respect, regardless of their identity or sexuality."*
Empathy/ Understanding (13 threads)	Express understanding of recipient's situation or disclose personal situation that communicates understanding	*"I understand how u feel. Moderating is a highly visible role, and suddenly many users know who you are, which can lead to some serious backlash."*
Encouragement (6 threads)	Provides recipient with hope and confidence	*"That's not okay. Don't give up. Last year, the subreddit I moderate was in even worse shape, but I persevered. You'll too."*
Jokes/memes (3 threads)	responding with joke or meme	*"Unofficial response from the admins (linked meme images)"*
Instrumental Support (38 threads)		
Tangible Resources (34 threads)	Share tangible resources, i.e., codes, tools, documents, etc. that can help to solve recipient's problem	*"Into the Automod, copy and paste this (Automod code). You will find 'Automod' option under mod tools"*
Willingness (8 threads)	Express willingness to perform tasks or provide services that would directly help with recipient's situation	*"Does my explanation make sense? I can look at it if you are worried about a particular element. "*

Table 3: Types of support received in the moderator support communities to manage community safety

Designing the Informing Process with Streamers and Bystanders in Live Streaming

Yanlai Wu
University of Central Florida

Xinning Gui
The Pennsylvania State University

Yuhan Luo
City University of Hong Kong

Yao Li
University of Central Florida

Abstract

The ubiquity of synchronous information disclosure technologies (e.g., live streaming) has heightened the risk of bystanders being unknowingly captured. While prior work has largely focused on solutions aimed only at informing the key stakeholder - bystanders, there remains a gap in understanding how device owners and bystanders mutually expect the informing process, which is critical to ensure successful informing. To address this gap, we utilized live streaming as a case study and conducted a design ideation study with 21 participants, including both streamers and bystanders. Our focus was to understand streamers' and bystanders' needs for informing regarding bystander privacy at the ideation state and derive design principles. Participants' design ideas reflected various and nuanced privacy concerns, from which we identified key design principles for future design.

1 Introduction

Synchronous information disclosure, where the creation and consumption of information occur simultaneously in the same space [19, 65], has surged in recent years. Live streaming, a popular example of this technology, facilitates real-time self-presentation, experience sharing, and interaction among users, exemplifying synchronous broadcasting [19, 59, 65]. Despite the benefits of immediate information sharing, synchronous information disclosure poses significant privacy risks to bystanders, who are inadvertently captured in device owners' information sharing [13, 15]. Bystanders, ranging from passersby to close contacts such as roommates and family members, often find themselves unexpectedly exposed, with limited control over the personal information they wish to keep private. The broad spectrum of personal data collected in synchronous information disclosure, including visual and auditory information, exacerbates the risks of exposure [5, 46, 64], thereby elevating privacy concerns [16, 31, 38]. Prior research and news have reported bystanders were worried about their personal information [28, 58] being captured

[11, 53], stalked [67], and misinterpreted [16, 58], leading to a series of negative consequences for them, such as financial loss [41], negative reputation [54], and harassment [7].

Prior research underscores the critical role of informing as a fundamental privacy protection for bystanders [2, 35, 60]. Informing here typically refers to enabling bystanders to be aware of the data sharing practices, the use of their personal information and the potential privacy risks [66]. Discussions have centered on enabling device owners to notify bystanders about their inclusion in information sharing activities [10, 28]. As the devices used for synchronous information disclosure like cameras and microphones become more integrated into everyday items like smartphones and wearables, it becomes increasingly difficult for bystanders to recognize when they are being recorded [2, 64]. Consequently, researchers have designed indicators [1, 2, 12, 16, 60], notifications [37, 46, 60, 68], and alerts [26, 52] to improve bystanders' awareness of potential privacy invasions. Yet, these solutions have largely designed based on the bystanders' perspective [1, 2, 37, 46, 60, 68], leaving the perceptions, concerns and challenges of device owners in the informing process largely unexamined.

On the other hand, the informing process should not be viewed as unidirectional (from device owners to bystanders only). Bystanders might also need to communicate their privacy preferences to device owners to ensure their privacy expectations are met. While this dynamic has been explored in the context of asynchronous information disclosure, such as allowing bystanders to express their consent and concerns to the photo owners [3, 55, 68], it has received less attention in synchronous settings. The instantaneous nature of synchronous information disclosure offers limited opportunity for bystanders to convey their preferences before being captured, raising questions about the mechanisms through which bystanders prefer to inform of their concerns and how device owners interpret and act upon such feedback.

Therefore, there exists a notable gap in research on reciprocal informing practices between device owners and bystanders. Several crucial questions remain unanswered: Do device owners intend to inform bystanders about their po-

tential inclusion? Do bystanders wish to communicate their privacy preferences to device owners? How do both parties view the informing process? What obstacles might they encounter while attempting to inform each other? To address these questions, it is essential to explore the mutual expectations of device owners and bystanders regarding informing.

Our study aims to fill this gap through a case study of live streaming. We chose live streaming for three reasons. First, live streaming is an increasingly popular form of real-time social media and it easily involves bystanders in various settings, such as private spaces, public areas, and online [16, 29, 30, 64]. Second, different from other synchronous devices such as IoT and AR, live streaming is more interpersonal as it allows direct and synchronous information disclosure to large anonymous viewers, which poses greater challenges for bystanders. Third, streamers, who are the device owners, aim to create content to attract a broad audience and earn profit [65], thus may weigh their own interests over bystander privacy. With live streaming as the study site, we seek to answer the following research questions:

RQ1: What needs, challenges and constraints of informing do streamers and bystanders have when it comes to bystander privacy in live streaming?

RQ2: What design do streamers and bystanders envision to address these needs, challenges and constraints?

To address these research questions, we engaged 21 participants with both streamers and bystanders in live streaming to conduct a design ideation study [22]. Design ideation involves generating, refining, and communicating ideas [25]. This process often marks the beginning of an imaginative and inventive approach [61]. In this paper, we used design ideation as a way to examine streamers and bystanders' various and nuanced design ideas that reveal both bystanders' and streamers' privacy needs for informing related to bystander privacy across streaming scenarios. Based on the ideas proposed by the participants and the design rationale for each idea, we derived design principles for the informing process that address bystander privacy.

The contributions of our paper are three-fold. First, we advance the understanding of bystander privacy by exploring the informing process from a multi-stakeholder perspective. Second, our findings reveal the multifaceted, nuanced, intricate, and dynamic nature of collective privacy management in the context of interpersonal and synchronized content-sharing platforms. Third, we propose key design principles for the informing through user-centric design, guiding the design of future synchronous information disclosure technologies.

2 Related Work on Informing Mechanisms for Bystander Privacy Protection

Extensive research has explored the privacy vulnerability of unaware bystanders in diverse socio-technical contexts
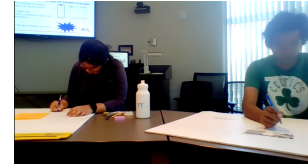
Figure 1: Two participants are designing in a design ideation study session

[2, 4, 16, 21]. These contexts involve both asynchronous and synchronous information disclosure. As bystander privacy was initially noted in asynchronous information disclosure (e.g, photo sharing) and then intensified with the proliferation of technologies in synchronous information disclosure (e.g., live streaming), we first review prior work regarding informing mechanisms in asynchronous information disclosure and next move to synchronous information disclosure.

2.1 Informing in Asynchronous Information Disclosure

In asynchronous information disclosure, users' content production and consumption do not happen at the same time and space [65]. Examples of asynchronous information disclosure are social media posts, instant messaging, discussion forums, and so on. In these contexts, users generate content that can be reviewed, edited, or withdrawn before being received by the recipients. Regarding bystander privacy in asynchronous information disclosure, most research attention is on collaborative photo sharing [10, 27, 34, 63]. Bystanders, who are pictured, mentioned, linked or tagged other than the photo owners who share the picture [10, 18], are often unaware of being featured in others' photo sharing [34, 68]. Bystanders in collaborative photo sharing lack control over their personal information, as photo owners decide how and with whom the images are shared [23, 24]. As such, bystanders often worry about appearing inappropriately, such as being captured as drunk or undressed in group photos on social media [28].

Researchers have proposed diverse informing solutions. Earlier research has focused on informing bystanders of bystanders' involvement [20]. Xu et al. [66] propose a face-recognition algorithm to identify and inform bystanders about potential privacy violations. More recent research shifts towards a more interactive approach, highlighting the necessity of obtaining consent from bystanders and facilitating negotiation between both two parties. For example, Facebook enables tagged bystanders to ask the photo owners to limit the visibility of their tagged photo [6]. Zheng et al. [68] propose an access-control protocol that mandates the consent of all parties in a photo before sharing it. Salehzadeh et al. [42] suggest a mediator to remind the photo owner to obtain consent from bystanders, and use middle-ground solutions to support conflict-solving [55]. Mosca & Such [40] present a multi-step

negotiation agent to discuss the sharing policy between both parties, considering their sharing preferences and moral values. Nourmohammadzadeh et al. [44] introduce a multi-agent system where an algorithm calculates user opinions based on user personality and behavior. However, the design solutions for asynchronous information disclosure may not apply to synchronous contexts where bystanders' data is collected in real-time, constantly, and goes beyond images.

2.2 Informing in Synchronous Information Disclosure

In synchronous information disclosure, users generate and consume information simultaneously [19, 65], such as IoT devices (e.g., smart home devices), wearable cameras, AR, and live streaming. Different from asynchronous information sharing, these technologies are normally less visible [1, 2, 64], operate continuously, and immediately capture a larger set of bystanders' personal information, such as bystanders' images, voices and movements [29, 37]. Due to this real-time, always-on, and continuous nature, it is more challenging for bystanders to be aware of being recorded in synchronous information disclosure than asynchronous information disclosure. This leads to bystander privacy concerns such as unknowingly being recorded saying inappropriate words [37], performing sensitive activities (e.g. withdrawing money on ATM) [53], or going to private locations [14].

To inform bystanders of their involvement in synchronous information disclosure, researchers have focused on design solutions across different technologies. In IoT, researchers aim to inform bystanders by enhancing the visibility and physical interaction with these devices. For example, Ahmad et al. [2] suggest that IoT devices should be designed to clearly display their sensor activities, such as showing on/off states to bystanders. Thakkar et al. [60] propose a bystander mode in mobile apps, enabling bystanders to view data relevant to themselves. Marky et al. [37] found that bystanders prefer various informing methods, including verbal communication, signs, notifications, and social media alerts. Pierce et al. [52] propose a mobile app that alerts bystanders about nearby smart cameras or microphones.

Prior work also indicates bystanders not only want to be informed but also want to have the option to control the IoT devices, which can offer them a sense of security [36, 49]. For example, Park et al. [49] propose different modes of control for bystanders in IoT. Bystanders in Marky et al.'s work express a desire to erase and block any data gathered about them by smart home devices [38]. Mare et al. [36] propose an interactive dashboard in smart home devices for Airbnb guests to access important details about and control the home's devices. Similar to this, Pierce et al. [52] introduce a guest account feature that allows bystanders limited access to IoT devices.

Besides IoT, for wearable devices, Perez et al. [50] developed FacePET, a wearable system for bystanders to manage privacy against unauthorized facial recognition. In AR, studies about informing are focused on informing AR users of the presence of bystanders to avoid interrupting the AR user's experience, rather than protecting the privacy of the bystanders [39, 45, 47].

In live streaming, to our best knowledge, only one study by Faklaris et al. [16] has proposed ways to inform bystanders in public and semi-public spaces, such as using colored lights on smartphones to signal active streaming and a 'Do Not Record' facial recognition database to blur registered faces. While [16] examines bystander attitudes toward being streamed in outdoor settings solely from the bystander's perspective, the informing process in live streaming involves multiple stakeholders, including both streamers and bystanders [64], the challenges and the designs envisioned in informing for multiple stakeholders remain unaddressed.

In sum, previous research related to synchronous information disclosure explores the informing process from the perspective of a single stakeholder, mostly the bystanders. Although Thakkar et al. [60]'s study includes both IoT device owners and bystanders, they examine each stakeholder's individual privacy needs rather than their mutual understanding of bystander privacy. However, bystander privacy protection requires mutual effort between the device owners and bystanders. As per communication privacy management (CPM) [51], all stakeholders need to negotiate and agree on their privacy expectations to ensure mutual privacy protection. While bystander is the key, the exclusion of device owners raises uncertainties about device owners' considerations in this matter. Therefore, it is crucial to design informing from a multi-stakeholder perspective.

Additionally, prior work focuses on synchronous technical platforms where bystander data are received by device owners or service providers to promote convenience [43] and automation [48]. However, with the rise of synchronous social platforms such as live streaming, information disclosure has become more interpersonal and involves social aspects such as self-presentation [65], relationship building [57], and interaction enhancement [9]. Thus, it is essential to integrate the interpersonal nature into the informing solutions.

To fill these gaps, we choose live streaming as a case study to explore the informing process about bystander privacy with bystanders and streamers. We conduct a design ideation study to delve into the multi-stakeholder perspectives and considerations, which will be debriefed next.

3 Methods

3.1 Participants & Recruitment

We conducted design ideation sessions (Figure 1) in April 2023 with a total of 21 participants, including 3 participants who identified themselves as streamers only, 8 as bystanders only, and 10 as both streamers and bystanders. This diverse

group was recruited to provide various viewpoints and develop comprehensive design solutions to address bystanders' privacy in live streaming. Our study was approved by the university IRB.

Our study's participants consisted of 14 males and 7 females, with ages ranging from 18 to 45 years old. Recruitment was conducted through various channels, including word-of-mouth, flyers, university mailing lists, and social media platforms (Facebook, Twitter, and Reddit). Participants who were interested in our study were first asked to complete an online screening survey. This survey included a consent sheet and questions regarding their demographics, how they want to participate in the research (in person or via Zoom), their email addresses, their role as a streamer and/or a bystander, and their experience with live streaming or being live streamed. Participants who are under 18 years old and have no experience as a streamer or bystander are excluded. The demographic information of participants is presented in Table B in Appendix.

According to participants' preferences and availability, we organized 3 online design ideation sessions through Zoom with an average of 2 participants per session, and 10 in-person design ideation sessions, with an average of 1 to 5 participants per session. While we aimed to have at least two participants in each session to foster diverse perspectives and collaborative brainstorming, some were unable to attend due to personal reasons, resulting in some solo sessions. To ensure a diverse range of perspectives, we strategically grouped participants in various combinations: bystanders only, streamers only, and bystander with streamer. To minimize potential biases, we ensured that participants who were acquainted with each other were assigned to different groups. Two researchers participated in the design ideation sessions. Each session took about 2 hours, and each participant was compensated with $40 in cash at the end of the study.

3.2 Design Ideation Sessions

Each session began with a warm-up activity, followed by a design ideation activity, and concluded with a debriefing interview. The primary goal of these sessions was to explore potential informing mechanisms to address bystander privacy concerns in live streaming.

Warm-up Activity Each session began with a round-table introduction. We then asked each participant to jot down 1-3 privacy challenges that bystanders would encounter in live streaming on a card. For online participants, they were asked to write down the challenges on their own papers. We displayed a slide to guide their thoughts. In the slide, we first explained who are considered bystanders in live streaming, including unknown passersby in public spaces, known people in the household (e.g., family, friends, and roommates), and virtual bystanders (e.g., in-game teammates and contacts in an online conversation). We then listed a set of questions, such as what personal information of bystanders was streamed,

any consequences, and where the disclosure happened. To avoid potential biases, we rephrased privacy in terms of personal information that bystanders do not want to share with the audience. Bystanders could draw from their own experiences, while streamers were encouraged to consider potential challenges their bystanders might face or as if they were bystanders streamed by others. After participants finished writing, we invited them to verbally share their thoughts. During this process, we asked follow-up questions to probe the details, such as whether the streamer had notified the bystanders, how the bystanders realized the streaming, how streamers realized their bystanders were being streamed, the relationship between streamers and bystanders, and the actions streamers or bystanders took afterward. This activity allowed participants to reflect deeply on bystanders' privacy challenges, thus setting a foundation for the design ideation activity.

Design Ideation Activity Following the warm-up activity, each participant was asked to design informing features to address the privacy challenges they mentioned in the last step. Our focus on 'informing' was driven by its importance to bystanders shown in prior work [2, 35, 60] and aligned with our research questions. In the slide prompt for this step, we told them they were free to design the features in any way they desired, without the need to consider existing technical constraints. We chose the word 'feature' for its broad applicability, encompassing both technical and non-technical solutions. Participants were encouraged to propose any type of design, including software, hardware, policies, procedures, etc. For our in-person participants, each participant was given a large flip chart paper (25" x 30") as a mock-up interface for a computer, phone, or hardware that was commonly used in live streaming. They were also provided with a set of paper-based design widgets (e.g., webcam, speaker, screen sharing, virtual background, overlay, beauty filter, chatbot, buttons, toggles) and craft supplies (glue, scissors, marker, tape, and sticky notes). These items were intended to spark creativity and provide a starting point for the participants' designs. Participants could use any provided widgets or modify existing widgets. They were also asked to annotate each design decision. Our online participants utilized Figma, an interactive online whiteboarding tool, to create informing solutions. We provided a brief tutorial on using Figma to ensure our online participants were comfortable with the tool. Each participant was provided with an individual Figma account to design so that they would not be affected by other's design. All the widgets were made digitally available on Figma. After the design, each participant was asked to present their informing design and explain the design rationale. We also asked questions to probe details such as how the design addressed bystanders' privacy challenges, who initiated the design, and the limitations of using the feature. Participants in individual sessions completed the activities individually. Participants in group sessions first worked independently and then shared with others, facilitating collaborative brainstorming.

De-briefing At the end of the study, each participant was asked to revisit and modify their designs. If they made any modifications, they would be asked to clarify the rationale for each revision. Participants also reflected on how their informing designs could protect bystanders from privacy violations in live streaming.

3.3 Data Analysis

We employed thematic analysis [8] with an inductive approach to analyze the data. Our dataset consisted of the video recordings of the design ideation sessions (totaling 19 hours) and paper/digital prototypes that participants created. These video recordings were transcribed into text, and the prototypes were digitized for analysis. Four researchers with domain expertise in live streaming and privacy research engaged in the data analysis process. Each researcher first independently examined the transcribed texts and the elements in the prototype and identified initial codes. We then compared and discussed our initial codes and combined them into a single list, resulting in 96 codes. Based on the codes in the list, we placed all the codes on Miro (an online whiteboard) to examine the relationships and patterns between codes, collate similar codes, and identify themes after extensive discussions. In this process, we continuously revisited the dataset and refined the themes and sub-themes. The final thematic map consists of two primary themes: the considerations of bystanders and streamers in the informing process, and the design ideas desired by both bystanders and streamers for informing.

4 Results

Our results showed that both bystander and streamer participants were concerned that when being streamed by others, bystanders would suffer from: 1) personal details such as location or phone number being exposed to viewers, leading to unauthorized physical and virtual contact, 2) harmful actions from malicious audiences, including ridiculing, doxing, swatting, and stalking, 3) financial information like credit card details being accidentally revealed through streamer's webcam, and 4) the streaming of unfavorable moments, such as having a bad hair or poor game performance. Our findings align with prior research on bystanders' privacy concerns in live streaming [16, 29, 30]. Therefore, we did not delve deeply into these concerns but rather briefly introduced them here to set the stage for the upcoming sections. To protect bystanders from privacy violations, both streamer and bystander participants expressed their desire to be informed, and they also hoped the other party to be informed. Streamers wish to know if bystanders want to be included in the stream, and bystanders want to be informed if they are or will be part of the stream. However, our participants reported various challenges in informing, which have not been discussed in previous work and we will discuss them next.
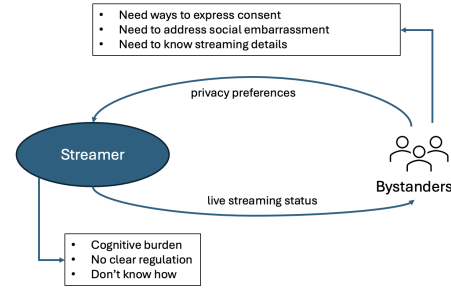
Figure 2: The bilateral communication informing loop

4.1 Needs, Challenges, and Constraints

Based on the interviews with streamers and bystanders, we found a bilateral communication loop in the informing process (Figure 2). Initially, streamers need to inform bystanders of the status of their live streaming. Once aware, bystanders then need to inform streamers of their privacy preferences, such as their willingness or concerns regarding being streamed. Bystanders (BS2[1], B5, B7, B11, B19, BS24, BS25) believed streamers should inform bystanders about the streaming status because bystanders often remain unaware of their inclusion in the stream, while streamers (BS6 and BS1) believed bystanders should inform streamers about their privacy preferences because streamers might not know bystanders' desire or reluctance to be streamed. However, there are challenges and constraints in this bilateral communication informing loop.

4.1.1 When Streamers Inform Bystanders of Their Live Streaming

Streamers have cognitive burdens. The real-time, multimodal, and socioeconomic nature of live streaming poses significant cognitive burdens on streamers to inform bystanders about the status of their live streaming. Our streamer participants (BS22, S23, BS25) agreed that as streamers, they needed to inform bystanders of their streaming activities. However, they were often deeply engaged in managing their performance, screensharing, audio/video input, and the synchronous interaction with their audience, to maintain a high quality of live streaming content. Given the real-time nature of their performance and interaction, streamers had to concentrate on the content they were broadcasting as they could not edit or withdraw the unwanted content. As such, streamers (S8, BS18, the streamers of BS4, BS6, BS24) sometimes forgot to notify bystanders in advance. As a result, our bystander participants (BS1, BS4, B11, B15, B19, BS24) shared that they were often informed after they had already been included in the stream, which made bystanders feel anxious. They often worried about whether they had said anything inappropriate

---

[1]"S" indicates the participant's self-reported role as a streamer, "B" indicates the participant's role as a bystander, and "BS" indicates that the participant is both a bystander and a streamer.

or done something 'stupid' (BS24) during the stream. Hence, streamers hoped they could be supported to inform the bystanders. For example, S8 said:

> I usually tell my parents I'm going to be busy, please don't come in during this time. But I think I forgot to let them know that day.

Informing bystanders is particularly burdensome in contexts involving large crowds such as public spaces. Streamers (S12, BS6) often found it burdensome and impractical to inform each bystander when they streamed in public settings because streamers were often preoccupied with their streaming activities and there were too many bystanders around to be informed. For example, S12 said:

> If in public, there's so many people around, I could see it being a burden if lots of people joining in and out and you tell them "hey, I'm streaming".

No clear regulations on informing. Our participants were uncertain about whether U.S. legislation permits or restricts live streaming in public spaces, and whether streamers are obligated to inform bystanders before their streams. The perceived lack of law clarity leads to varying opinions about whether informing should be given in public spaces. Some streamers (BS22, S23) and bystanders (BS2, B5, B16, BS25) believed that informing should occur regardless of lacking strict regulations for streaming in public spaces. These streamers acknowledged an ethical responsibility not to stream people randomly. These bystanders felt that streaming in public without informing them violated their portrait and privacy rights. However, some streamers (BS6) and even bystanders (BS1, BS24) felt streamers had the right to stream in public spaces and bystanders were not obligated to be notified because there were no regulation requirements. These different opinions indicate that clear regulations, policies and guidelines should be specified regarding informing bystanders. For example, BS1 (from bystander perspective) reported:

> I just had a quick question. I don't know if there's any legal framework behind streaming in public. Do you guys know anything? [...] But I feel like in public spaces, I think it's usually fine for you to just stream yourself. I don't think there's any legal restriction. I don't think there should be something that should be notified. I don't think I'm obligated to be notified.

Streamers struggle with how to inform bystanders. In most cases, streamers (S12, BS6, S23) did not know the bystanders. For instance, bystanders in public and online spaces are mostly strangers or passersby. Even when streamers wanted to inform their bystanders, they typically lacked the means to contact the bystanders, such as bystanders' phone numbers or social media accounts. For example, BS6 (from streamer perspective) reported:

> I'd tell them I'm streaming probably through a message if I had their number, but that hasn't always been the case.

Furthermore, our participants (BS1, BS2, B5, BS6, B7, B11, B19, BS24, BS25) reported that bystanders often lacked direct access to the streaming platform, making it impossible for streamers to reach each bystander within the streaming platform. For example, BS24 (from streamer perspective) said:

> Grant bystanders direct access to the streaming is challenging, since it would require them to log in as co-hosts, which goes against the intention of those who don't plan to be on the stream.

4.1.2 When Bystanders Inform Streamers of Their Privacy Preferences

After bystanders become aware of the potential to be exposed in a live stream, they need to inform streamers of their privacy preferences. During the process, they have to navigate through various challenges and require additional information to effectively communicate their preferences.

Bystanders need ways to clearly express their consent. Bystander participants (B5, B19, BS24, BS25) sometimes prefer to seek a more interactive consent process especially before they are captured in others' live streaming. They wanted to explicitly express their agreement or disagreement with being streamed to the streamers. With the consent, bystanders felt a sense of respect and might be more willing to be part of the streaming. Streamers could also realize bystanders' willingness or not and more effectively protect bystander privacy. However, bystanders often have limited ways to explicitly express their consent to streamers, unless the streamers intentionally ask for their consent. For example, B5 said:

> I've been streamed as focus, I was being asked pop-up quiz on campus, but that was with my consent. They approached me and asked me, 'would you want to be a part of this?' I said, 'sure, why not?', they asked me for my consent. This one I was asked and I said ok actually, so it was okay.

Bystanders need to cope with social embarrassment. Even when bystanders are approached to express their privacy preferences, they (BS2, B3, B5, B7, B11, BS25) frequently hesitated to explicitly communicate their privacy preferences with streamers due to social embarrassment. This hesitation was often rooted in politeness and a belief that they were not the primary focus, particularly when the streamers were unknown to them. When the streamer was known, bystanders also worried that such direct conversation might negatively influence the relationship between the known streamer and the bystander. Even when bystanders wanted to explicitly express their unwillingness, bystanders believed the streamers might misinterpret their concerns. As such, bystanders

would choose implicit actions, such as dodging the camera and walking away, rather than directly talking to the streamer about their concerns. Some participants (BS1, BS6, B11) even favored merely informing without a clearly indicated consent to avoid direct interaction with streamers. Bystanders found a sense of "peace of mind" when they did not have to provide explicit consent. For example, B7 told us:

> What I'll do is to say, 'Hey, I really didn't like this'. Sometimes, they'll respond with, 'it's not that serious', 'don't take it so serious' or 'they didn't have any malicious intent'. They feel like I come after them and that's not what I'm trying to do.

However, streamer participants (BS6 and BS1) argued that without explicit communication, streamers might not know bystanders' reluctance to be streamed, nor take actions, such as adjusting the camera angle or relocating to a different area to avoid bystanders being streamed. Therefore, there needs an approach to deliver bystanders' privacy preferences to streamers without causing social embarrassment.

Bystanders need to know how they are streamed. Bystanders need details to make informed decisions on whether to be captured in others' live streaming. Streamers often did not provide detailed information about their streaming to bystanders, believing that bystanders were either not familiar with the concept of streaming or did not have access to the streaming platform that they use. However, bystanders (B3, BS4, B11, BS18, B19, BS25) expected more in-depth explanations, such as the specific streaming platforms being used, the devices being enabled, the intended audience, the live stream's topic, and the reasons for their inclusion as bystanders. Such detailed information was crucial for bystanders as it helped them to assess the potential reach of the stream and to understand how their presence might be interpreted or utilized in the stream, which are key to their privacy decisions. For instance, in private space, bystanders often have close proximity to the streamer's webcam and microphone, increasing the chance of their accidental appearance and voice capture in streams. So they wanted to know whether and how their appearance and voice could be captured. For example, BS18 (from bystander perspective) told us:

> I didn't know my roommate was streaming, and we shared a room. So I was just back from the gym and like the way his camera is set up, he can see like, the whole room is visible. So no matter when I come in or go out, he can see and everyone else can see it too. It'd be a good idea if he could just send a snapshot of how the webcam is placed and what it's capturing at the start of the stream.

This is also the case in online space. Streamers often stream on multiple streaming platforms to reach a wider audience and earn more money. As a result, bystanders are streamed on

multiple platforms, thereby amplifying their exposure and privacy risks. Hence, bystanders need to know all the platform(s) where they are live streamed. For example, BS4 (from the bystander perspective) said:

> It'd be nice to know the streaming platform, cause if it's just streaming discord, I know it's just like three people watching, if on Twitch, it might be 300 people watching.

4.2 Desired Design Solutions

Given the needs, challenges and constraints reported in the above section, participants suggested various design solutions for the informing process to facilitate streamers to inform bystanders of their live streaming activities and also facilitate bystanders to inform streamers of their privacy preferences. In this section, we will introduce these design ideas.

4.2.1 Platforms-Initiated Automatic Alerts

To reduce streamer's cognitive burden and minimize streamer's effort in informing bystanders, participants suggested that the live streaming platforms enable two types of automatic alerts for both streamers and bystanders:

Alerting streamers of bystanders' involvement. BS6 introduces an automated alert system for live streaming platforms designed to promptly inform streamers when a bystander is detected in their live streams within the physical environment. This system identifies bystanders via behavior or speech recognition techniques during the stream. Upon detection, streamers are alerted through a pop-up notification on their streaming device, highlighting the presence of a bystander. They can then take actions, such as informing the bystanders or avoiding bystanders being streamed. This feature requires no effort from bystanders and aims to make streamers aware of potential privacy violations to bystanders, particularly when streamers are busy with their live performances. With this automated alert, streamers can remain focused on their live performance, alleviating the need to constantly monitor for the involvement of bystanders. For example, BS6 explained based on his streamer experience:

> If I was actively playing a game, then it's hard to realize there is a bystander in that case. If I had to send them notifications all individually, I think that would probably get too complex to manage. One option could be to have the streaming platform do that detection for you. And if it notices certain behavior or certain words being said, it could pop up a notification saying like 'someone getting in'.

Alerting bystanders before streaming. Our participants introduced an automatic alert feature designed to notify bystanders within the virtual environment when streamers start live streaming. This system requires bystanders to proactively

a) Provide gamers a link to the streamers' streaming channel

b) Send bystanders a prerecorded message along with a snapshot of the streaming area

c) Ask the streamer to wear t-shirts about the streaming

d) Bystanders register in a opt-out database to be auto-blurred when they enter the streaming

e) Bystanders receive an automated SMS notification, and their consent responses are aggregated into a notification for the streamer.

Figure 3: Examples of participants' design. We used Canva (an online protyping platform: https://www.canva.com/) to digitally translate the paper-based sketches created by the participants.

follow streamers, such as their roommates, gaming partners, and friends, who might inadvertently capture and broadcast their personal information during live streams. Once these followed streamers initiate a live stream, the bystanders immediately receive a notification from the platform, ensuring they are aware of the streaming activity without relying on the streamers to inform them manually, who might be preoccupied with their performance. Streamers are only required to guide their potential bystanders on how to follow their streaming accounts. With this setup, every time streamers go live, the platform automatically alerts the bystanders, allowing streamers to focus on their content without the need to individually notify each bystander. For example, B10 suggested:

> You can follow people on Twitch. This feature is useful if your friends in the game are also your friends on Twitch. For instance, if I follow my buddy on Twitch and he starts streaming, I would get a notification on my phone through Twitch, alerting me, 'Hey, your buddy is streaming.' This happens even if he doesn't directly tell me.

4.2.2 Platform-Enforced Regulation and Policy

To mitigate the uncertainty surrounding legal and policy requirements about streamers' obligations to notify bystanders who are within the physical environment before streaming them, participants recommended the establishment of explicit policies. These policies could be enforced through legislation or platform guidelines, such as terms of conduct, community guidelines, or tutorials for new streamers. The policies should clarify whether streamers are required to inform bystanders and obtain explicit consent from bystanders, especially in public spaces. This approach aims to legally protect bystanders' privacy and ensure ethical streaming practices by clearly defining the informing responsibilities of streamers towards bystanders. For example, BS2 (from bystander perspective) said:

> Maybe like a policy the streamer has to agree to. You have to get consent before you stream anybody or something.

4.2.3 Embedded Communication Channels to Inform

As streamers reported challenges in delivering their informing to the bystanders, and bystanders also struggled with ways to express their consent, participants designed a series of communication channels embedded or linked with the live streaming platforms to assist with the informing:

One-to-one messaging between streamers and bystanders. Participants suggested implementing a two-way one-on-one messaging feature within live streaming platforms. This feature would facilitate communication between streamers and bystanders who are within the physical environment in two main scenarios: informing bystanders about live streaming activities and allowing bystanders to convey their privacy preferences to streamers. When streamers plan to go live and wish to notify bystanders, they can request the platform to send a notification message directly to those bystanders. If the bystanders are registered users of the platform, BS17 proposed that bystanders could receive this notification through an in-platform message, enabling streamers to inform them without needing to access their private contact information. Bystanders can then respond within the same platform, stating their privacy preferences clearly and directly:

> I share kitchen space with my roommate. I came into the kitchen to make myself some stuff to eat. I had no idea she was streaming (B15's concern)

> Her roommate can tag her [...] Similar to Facebook tagging or identity of that person and then send the person a notification. (BS17's design)

For bystanders not registered with the platform, i.e., they have not installed the app or are unfamiliar with live streaming, B5 proposed a GPS-based SMS feature (Figure 3(e)) to provide a communication channel between streamers and bystanders. Streamers would first input their streaming location. Bystanders who are nearby would then receive an automated SMS notification stating, "You are entering a streaming vicinity, do you want to be streamed?". This approach necessitates collaboration between platforms and government or service providers to send automated SMS alerts to bystanders. Such collaboration could effectively communicate the consent process to a broad audience swiftly and directly, bypassing the

need for streamers and bystanders to exchange contact information. The system also allows bystanders to respond with their consent. These responses could be aggregated into an average approval rating for the area, which will serve as a guide for streamers to gauge general bystander consent:

> I thought of a feature, which is like a GPS-based authentication thing for crowd streaming [...] We all receive some SMS alerts when the hurricane or kidnapping happens, because we are in this sort of area which was impacted by it. We could approach the government and don't have to ask anybody for their number, you just send them an alert, they can choose whether they approve of been seen or not [...] Definitely it's not possible to go and ask everybody for the approval, but we'll take an average. If the on average approval rating is quite low, then the streamer must probably reconsider doing it somewhere else.

Both one-on-one messaging functions eliminate the need for streamers and bystanders to collect each other's personal contact information, streamlining the process of informing. It also offers bystanders a straightforward method to articulate their privacy preferences, ensuring they have a say in whether they want to be included in live streams. Additionally, the one-on-one messaging feature guarantees that notifications are delivered and seen, even if streamers or bystanders are otherwise occupied, thereby enhancing the effectiveness of communication between streamers and bystanders.

One-way one-to-many indicator. Aside from the messaging communication channels, participants also suggested one-way one-to-many channels to inform bystanders through visual or auditory indicators. Such indicators can be physical or virtual indicators initiated by streamers. Once activated, it will notify broad bystanders about the streaming status through visual or auditory cues. The indicator is independent of sending individual notifications to bystanders, as in public settings and online gaming scenarios, it is ineffective to individually message a substantial number of passersby and online players who keep coming and going in others' live streams.

The physical indicators include visual cues, such as flashing lights (S17, BS18) and conspicuous signs (B5), alongside auditory signals, such as a beep sound (S17). The indicators could be incorporated as a software feature enforced by the streaming platform or activated manually through a button on streaming devices by streamers when the streaming starts. They might even be integrated as a sign on the T-shirt (Figure 3(c)). The virtual indicators can be a "live" icon displayed around the game avatar of the streamers who are streaming. The icons are mandatory before streamers start their streaming online. These indicators help streamers inform the bystanders of live broadcasts without the need to obtain bystanders' contact information. For example, B5 reported:

> I think that's a good start to make streaming obvious because if it's a busy place, you can't really go and notify everybody and people keep coming and going. Maybe ask the streamer to wear t-shirts about the streaming.

However, our participants also acknowledged that the effectiveness of physical indicators could be influenced by the surrounding environment. Visual cues, for example, can become less effective in well-lit daytime environments where they might blend into the background, making it difficult for bystanders to notice them. Furthermore, there are concerns raised by some participants regarding auditory signals, which they find to be annoying. For example, B2 said:

> Because like you said, the blinking lights, right? But if it's daytime, right? Sunny, how are you going to see the lights? So stuff like that. That's why I said like, it's kind of difficult to implement it. And that's what I mean.

This highlights the necessity of designing indicators that are less intrusive and consistently recognizable, regardless of the environmental conditions they are subjected to.

4.2.4 Embarrassment-Free Bystander Privacy Expression

To mitigate the social discomfort bystanders may experience when expressing their privacy preferences not to be included in others' live streams, two embarrassment-free mechanisms for bystanders' privacy preference expression were proposed. **Bystanders' one-sided opt-out.** To accommodate those who wish to remain outside the scope of live streams, an opt-out mechanism (Figure 3(d)) has been proposed by our participants. This mechanism is designed for bystanders who are within the physical environment. Bystanders can register in a database provided by the platform or a third party, entering specific personal details that can recognize their identities. Upon registration, the system is designed to recognize these bystanders whenever they appear in others' live streams, automatically applying blurring or muting effects to their likeness or voice. This approach eliminates the need for any direct communication between bystanders and streamers, thereby sparing bystanders from the potential awkwardness of confronting streamers about their privacy preferences. For example, BS6 (from streamer perspective) reported:

> There's also the idea of on a platform-wide level, maintaining an opt-out type database of people who don't want to be seen on any stream ever. So what you could do is have any stream search through that database, but constantly cross-reference between the people it sees in the stream. If someone shows up that's in the database, automatically blur them out and mute them.

Device-enforced consent-based protection. To further alleviate the discomfort of bystanders in expressing their privacy preferences directly to the streamers, a device-enforced consent-based mechanism is proposed for bystanders who are within the physical environment. This feature automates the blurring of bystanders in live streams based on their consent, operating at the device level. When a streaming device, such as a camera or a microphone, detects a bystander, it triggers a notification to the bystander requesting their consent to be included in the stream. Should the bystander agree, they will appear unaltered; if they decline, the device will automatically blur their presence in the live stream. This method addresses bystanders' concerns that voicing their reluctance to participate might be perceived as impolite, lead to social embarrassment, or be disregarded by the streamers. By sending the consent to the device, rather than to the streamer, bystanders can assert their preferences without fear of personal conflict or judgment. The device-level enforcement ensures that bystanders' privacy preferences are respected as they are, granting them greater control and reducing the likelihood of misunderstanding. For example, B19 expressed:

> Streamer starts the live stream, and then people around the person where the camera can see clearly get an alert, like you have entered a live streaming vicinity. And then the bystanders can be asked if they are willing to join, because there are sometimes bystanders want to join. There are some people, sometimes people like being on camera, like being a part of something, if they like the streaming topic. And depending on that, they will be blurred out or no. If they want to be in the live stream, then they don't need to be blurred out.

4.2.5 Providing Details in Streamer's Informing

Bystanders require comprehensive details about the live streams in which they are to be captured to make informed decisions regarding their privacy preferences. They have voiced a need for detailed information on how, where, and why they are streamed. Irrespective of the method used by streamers to inform them — be it through a message, indicator, or alert — bystanders wish to be informed about the particular streaming platforms in use, the devices utilized for streaming, the target audience, the subject matter of the live stream, and the rationale behind their inclusion as bystanders. This information is crucial for them to make informed consent decisions. For example, in the case of a virtual indicator, bystanders expect it to provide a link to the streaming channel (Figure 3(a)). This link should clearly indicate whether the streaming is occurring on the same platform, a different one, or across multiple platforms. The link serves not only to give online bystanders an easy way to see how they appear in the stream but also to understand the potential audience size. For instance, BS4 (from bystander perspective) proposed:

[...] Right now, like discord, only sees if you're streaming on discord. You could join discord and not know if your friend is streaming on Twitch. The design I have was like a streamer mode for streamers. They have to enable the streamer mode to go live [...] For example, User1 is my friend. He's streaming on Twitch. I load up discord. It would show me he's in the voice chat. He's streaming on a different platform. Streamer could also give you the link of their channel. If you click on it, you could see the streaming status. I think that would be a good idea. I'm sure streamers would like that too, because then people can click and they need more audience.

If the notification comes as a message, our bystander participants within the physical environment recommend it include a pre-recorded message saying, "I'm streaming," along with a snapshot of the streaming area (Figure 3(b)). This feature allows bystanders to actively avoid areas where they might be captured on camera if they prefer not to be included in the live stream. It empowers them to navigate their environment with greater confidence and without the constant fear of unintentionally appearing in a live stream. For example, BS18 (from bystander perspective) said:

> I didn't know my roommate was streaming, and I just went into the room. So there might be a pre-recorded message like 'I'm streaming in my room' sent by a toggle bar. When they start streaming, they just send a snapshot of how the webcam is placed and what it's capturing. You can just avoid that area in particular, and just do everything that you want to do, and it's not being that area.

5 Discussion

Our study explores the reciprocal informing practices between streamers and bystanders in live streaming from both perspectives. Based on participants' design ideas, we propose three key design principles to enhance bilateral informing interaction in synchronous information disclosure.

Contextualizing informing process. Contextualizing is not a novel concept. Previous research has shown that privacy notices and choices should be tailored to contextual factors such as space, timing, channel, and modality [17,56]. This has been supported by prior work in synchronous information disclosure like IoT and AR, which considers contextual factors like environments (e.g., home, Airbnb) and pre-existing relationships between the owner and bystanders (e.g., host and guest) [2,37,52,60]. Our findings in live streaming also highlight the importance of environments and relationships for informing. For example, when device owners and bystanders are unknown to each other, participants prefer one-to-many

indicators in public spaces. In private spaces, with known contacts, one-on-one messages are preferred. In online spaces, where bystanders might be streamed on different streaming platforms, notices could be sent across various platforms.

Our contribution extends beyond prior work by emphasizing the **social and interpersonal nature** when contextualizing the informing process. The informing processes should adapt to the social contexts, considering factors such as the target audience, group size, and activities or status of stakeholders. For example, streamers might interact with a small, known group of viewers or a large, anonymous public audience. In cases of large and unknown audiences, it is crucial for bystanders to be informed about their visibility and data sharing. Additionally, the current activity and status of stakeholders, especially device owners, should be considered. Our findings indicate that streamers have to manage multiple activities, including performing, interacting with numbers of unknown audiences, and addressing context changes, which leads to a significant cognitive burden that may hinder their ability to notice and protect bystanders effectively. Thus, automated notifications that consider social aspects are necessary.

These contextual factors could also implicate other synchronous information disclosure contexts beyond live streaming. For example, bystanders might be involved when IoT users share recordings from their smart cameras with their friends, family members, and online social networks, thus expanding bystanders' exposure to different types of audiences. AR users might be preoccupied with interacting with other users and may not easily realize bystanders' unwillingness when bystanders pass by the device. Therefore, our design principles could contribute to other synchronous scenarios. We recommend that designers consider both the informing mechanisms for different contexts and the social and interpersonal nature of the informing process.

Balancing the power dynamic between streamer and bystander through mutual transparency. Our results highlight that the power dynamic between streamers and bystanders is unbalanced. Although there are bilateral informing needs, streamers need to initiate the informing process as bystanders often lack agency over their privacy. Bystanders rely on streamers for information and decision-making. Without streamer initiation, bystanders often lack awareness of being streamed and cannot make informed decisions. They expressed a desire for greater transparency, such as knowing the number of audience members, their presentation in the streams, and the streamers' attitudes toward their participation. Moreover, without explicit communication from bystanders about their privacy preferences, streamers remain unaware of bystanders' desires to participate or their level of comfort, exacerbating the power imbalance. Therefore, mutual transparency in the informing process is crucial, enabling both parties to make informed decisions and protect bystander privacy effectively.

To empower bystanders, previous informing designs in synchronous information disclosure, such as IoT, have enhanced transparency by detailing which devices are collecting data, what data is being collected, and whether data collection is active [2, 37]. Our research aligns with these findings. For instance, our bystanders also wanted to know if streaming is active and whether the streamer is using a camera or engaging in voice chat.

Our unique contribution emphasizes **mutual transparency** to reduce the power imbalance between bystanders and streamers by ensuring both parties have agency and are informed about crucial details, such as platform or legal policies, data recipients, social implications, and participation details. For instance, informing both bystanders and streamers about the platform or legal policies is especially important on online platforms and in public spaces where clear informing practices are often lacking. Without clear regulations, both parties may be unaware of bystanders' rights and streamers' obligations. It is also important to inform bystanders about data recipients, as live streaming involves broad, nontransparent, anonymous, and public audiences. Bystanders need to know who is receiving their data and on which platform. Additionally, providing information about social implications allows bystanders to manage their self-presentation to the audiences; for example, one bystander wanted to know how the streamer's audience commented on him. Lastly, informing streamers about bystanders' detailed participation information helps streamers understand if bystanders are willing to be part of the streaming, how much they want to participate, and in what streaming topics they want to be involved. These transparency details can also benefit other synchronous information disclosure contexts. For example, in the case of wearable health devices used in fitness centers, it is crucial to inform device owners and bystanders about who has access to the collected health data, the legal policies governing its use, the social implications of wearing such devices in public, and whether bystanders are comfortable being recorded or included in data collection.

While enhancing these details of the informing process, it is vital to consider that some bystanders may lack access to or interest in the streaming platform, limiting their control over their privacy. Therefore, informing practices should be transparent and effective without burdening bystanders, such as through visible indicators (i.e., one-way one-to-many indicators) or familiar communication methods (i.e., text messages) that do not require downloading an extra streaming platform. This consideration also contributes to other synchronous information disclosure contexts. For example, when a bystander is involved in AR interactions by AR users, bystanders might not be able to gain access to the AR device. In this case, informing practices should utilize methods that are easily accessible and do not impose additional steps on bystanders, ensuring their awareness and consent without requiring direct interaction with the AR device.

Mediating communication barriers between streamer

and bystander. Our findings reveal that the informing design should mediate the communication barriers between streamers and bystanders, especially through third parties such as platforms or government agencies. Our streamers want bystanders to inform them of bystanders' privacy preferences, but bystanders often feel embarrassed to confront streamers directly or do not trust the streamers' decisions in protecting bystander privacy. Our bystanders expect streamers to respect their privacy, but streamers might not do so because they assume bystanders are fine with being streamed or make decisions on behalf of bystanders at that point. Thus, **third-party mediation**, such as device or platform-enforced informing designs, can play important roles in mediating the communication barriers between streamers and bystanders.

Prior informing designs in synchronous information disclosure such as IoT have enabled bystanders to communicate with device owners about their privacy preferences or to circumvent device owners to make their own privacy decisions through mechanisms such as bystander mode [60] or guest accounts [37]. But do bystanders prefer to use these controls? Our findings in live streaming highlighted that bystanders sometimes do not prefer direct communication or control because bystanders felt it was bothersome to take extra steps to communicate with streamers, especially when they do not have easy access to the device. They also worried that such actions might be interpreted as impolite by unknown streamers or negatively influence their relationship with known streamers. However, they also do not want streamers' personal decisions or adjustments to override or influence their willingness to participate or be involved. Therefore, our participants proposed device/platform-enforced informing mechanisms such as opt-out database, platform-initiate alerts, platform-enforced policy, and device-enforced consent-based protection to provide assurance and fairness for bystander privacy without relying on streamers' subjective decisions.

Prior informing work in live streaming [16] also proposed using an opt-out database, but researchers proposed it from a one-sided perspective by providing bystanders with low-effort notifications. In contrast, our participants developed the opt-out database to address the communication barriers between the two stakeholders from a two-way perspective. Specifically, it is designed by our streamers to tackle bystanders' social embarrassment when expressing unwillingness to be streamed. It involves streamers actively recognizing and respecting bystanders' privacy preferences while also allowing bystanders to communicate preferences without direct confrontation. This approach demonstrates mutual privacy consideration, showing that one stakeholder group sincerely values the privacy and social needs of the other, emphasizing that bystander privacy protection requires collaboration among stakeholders. It highlights promising opportunities for cooperation and coordination between stakeholders in live streaming. Such two-way third-party mediators also have implications for other synchronous information disclosure

contexts. For example, bystanders might not feel comfortable directly communicating with AR users about their privacy concerns; thus, they could register their preferences not to be recorded, and the AR system would automatically blur their image or mute their voice, respecting their privacy without direct confrontation.

6 Limitation

First, our study concentrated on the ideation phase to foster innovative design ideas to address informing challenges regarding bystander privacy in live streaming. However, it did not include subsequent stages, such as prototype development and evaluation, which could have provided practical and validated design solutions. Future studies may implement and test the proposed ideas if technological advances allow.

Second, although we aimed to include at least two participants per sessions, representing both streamers and bystanders, but unforeseen absences led to some sessions with a single participant. While designing with one participant is common in prior work [32, 33], and can provide detailed individual insights [62]. But the varying group sizes may have affected the ideation outcomes and limited the diversity of of perspectives. Future work could aim to standardize group sizes to ensure more consistent and comprehensive insights.

Third, despite efforts to recruit participants through various platforms, most of our participants were college students. This might be because our study was conducted at the university. People with different professions or educational backgrounds may have different perceptions and practices related to managing bystanders' privacy. Therefore, our sample may not fully capture the perspectives of streamers and bystanders with different occupations or educational backgrounds.

Fourth, although we targeted participants from various disciplines, we had more CS students (43%) than non-CS (33%), with 5 participants not disclosing their majors. Since CS students tend to be more tech-savvy, our results might not accurately reflect the privacy needs of non-CS users. Future research could include participants from more diverse backgrounds to broaden the applicability of our findings.

7 Conclusion

In this paper, we engaged 21 streamers and bystanders to understand their mutual expectations for the informing process regarding bystander privacy in live streaming. The results suggested that both streamers and bystanders face a variety of challenges during the informing process in live streaming. Based on these insights, our participants proposed various design ideas for informing streamers and bystanders to protect bystander privacy. From these concepts, we summarized key design principles that can guide the development of future technologies in this area.

References

[1] Imtiaz Ahmad, Taslima Akter, Zachary Buher, Rosta Farzan, Apu Kapadia, and Adam J Lee. Tangible privacy for smart voice assistants: Bystanders' perceptions of physical device controls. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–31, 2022.

[2] Imtiaz Ahmad, Rosta Farzan, Apu Kapadia, and Adam J Lee. Tangible privacy: Towards user-centric sensor designs for bystander privacy. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–28, 2020.

[3] Gulsum Akkuzu, Benjamin Aziz, and Mo Adda. Towards consensus-based group decision making for co-owned data sharing in online social networks. *IEEE Access*, 8:91311–91325, 2020.

[4] Rawan Alharbi, Mariam Tolba, Lucia C Petito, Josiah Hester, and Nabil Alshurafa. To mask or not to mask? balancing privacy with visual confirmation utility in activity-oriented wearable cameras. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 3(3):1–29, 2019.

[5] Julia Bernd, Ruba Abu-Salma, and Alisa Frik. {Bystanders'} privacy: The perspectives of nannies on smart home surveillance. In *10th USENIX Workshop on Free and Open Communications on the Internet (FOCI 20)*, 2020.

[6] Andrew Besmer and Heather Richter Lipford. Moving beyond untagging: photo privacy in a tagged world. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1563–1572, 2010.

[7] Bitdefender. What are Private Data Leaks. https://www.bitdefender.com/cyberpedia/what-are-private-data-leaks, 2022.

[8] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

[9] Chia-Chen Chen and Yi-Chen Lin. What drives livestream usage intention? the perspectives of flow, entertainment, social interaction, and endorsement. *Telematics and Informatics*, 35(1):293–303, 2018.

[10] Hichang Cho and Anna Filippova. Networked privacy management in facebook: A mixed-methods and multinational study. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 503–514, 2016.

[11] Soumyadeb Chowdhury, Md Sadek Ferdous, and Joemon M Jose. Bystander privacy in lifelogging. In *Proceedings of the 30th International BCS Human Computer Interaction Conference 30*, pages 1–3, 2016.

[12] Camille Cobb, Sruti Bhagavatula, Kalil Anderson Garrett, Alison Hoffman, Varun Rao, and Lujo Bauer. "i would have to evaluate their objections": Privacy tensions between smart home device owners and incidental users. *Proceedings on Privacy Enhancing Technologies*, 2021.

[13] Matthew Corbett, Brendan David-John, Jiacheng Shang, Y Charlie Hu, and Bo Ji. Bystandar: Protecting bystander visual data in augmented reality systems. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*, pages 370–382, 2023.

[14] Tamara Denning, Zakariya Dehlawi, and Tadayoshi Kohno. In situ with bystanders of augmented reality glasses: Perspectives on recording and privacy-mediating technologies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2377–2386, 2014.

[15] Mariella Dimiccoli, Juan Marín, and Edison Thomaz. Mitigating bystander privacy concerns in egocentric activity recognition with deep learning and intentional image degradation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(4):1–18, 2018.

[16] Cori Faklaris, Francesco Cafaro, Asa Blevins, Matthew A O'Haver, and Neha Singhal. A snapshot of bystander attitudes about mobile live-streaming video in public settings. In *Informatics*, volume 7, page 10. MDPI, 2020.

[17] Yuanyuan Feng, Yaxing Yao, and Norman Sadeh. A design space for privacy choices: Towards meaningful privacy control in the internet of things. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2021.

[18] Ricard L Fogues, Pradeep K Murukannaiah, Jose M Such, and Munindar P Singh. Sharing policies in multiuser privacy scenarios: Incorporating context, preferences, and arguments in decision making. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 24(1):1–29, 2017.

[19] Johann N Giertz, Welf H Weiger, Maria Törhönen, and Juho Hamari. Content versus community focus in live streaming services: How to drive engagement in synchronous social media. *Journal of Service Management*, 33(1):33–58, 2022.

[20] Rakibul Hasan. Reducing privacy risks in the context of sharing photos online. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2020.

[21] Rakibul Hasan, David Crandall, Mario Fritz, and Apu Kapadia. Automatically detecting bystanders in photos to reduce privacy risks. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 318–335. IEEE, 2020.

[22] Noe Vargas Hernandez, Jami J Shah, and Steven M Smith. Understanding design ideation mechanisms through multilevel aligned empirical studies. *Design studies*, 31(4):382–410, 2010.

[23] Hongxin Hu, Gail-Joon Ahn, and Jan Jorgensen. Detecting and resolving privacy conflicts for collaborative data sharing in online social networks. In *Proceedings of the 27th Annual Computer Security Applications Conference*, pages 103–112, 2011.

[24] Haiyan Jia and Heng Xu. Autonomous and interdependent: Collaborative privacy management on social networking sites. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4286–4297, 2016.

[25] Ben Jonson. Design ideation: the conceptual sketch in the digital age. *Design studies*, 26(6):613–624, 2005.

[26] Marion Koelle, Katrin Wolf, and Susanne Boll. Beyond led status lights-design requirements of privacy notices for body-worn cameras. In *Proceedings of the Twelfth International Conference on Tangible, Embedded, and Embodied Interaction*, pages 177–187, 2018.

[27] Jacob Kramer-Duffield. *Beliefs and uses of tagging among undergraduates*. The University of North Carolina at Chapel Hill, 2010.

[28] Airi Lampinen, Vilma Lehtinen, Asko Lehmuskallio, and Sakari Tamminen. We're in it together: interpersonal management of disclosure in social network services. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 3217–3226, 2011.

[29] Yao Li, Yubo Kou, Je Seok Lee, and Alfred Kobsa. Tell me before you stream me: Managing information disclosure in video game live streaming. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–18, 2018.

[30] Zhicong Lu, Michelle Annett, and Daniel Wigdor. Vicariously experiencing it all without going outside: A study of outdoor livestreaming in china. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–28, 2019.

[31] Zhicong Lu, Haijun Xia, Seongkook Heo, and Daniel Wigdor. You watch, you give, and you engage: a study of live streaming practices in china. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–13, 2018.

[32] Yuhan Luo, Peiyi Liu, and Eun Kyoung Choe. Co-designing food trackers with dietitians: Identifying design opportunities for food tracker customization. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019.

[33] Stephann Makri, Tsui-Ling Hsueh, and Sara Jones. Ideation as an intellectual information acquisition and use context: Investigating game designers' information-based ideation behavior. *Journal of the Association for Information Science and Technology*, 70(8):775–787, 2019.

[34] Ameera Mansour and Helena Francke. Collective privacy management practices: A study of privacy strategies and risks in a private facebook group. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–27, 2021.

[35] Shady Mansour, Pascal Knierim, Joseph O'Hagan, Florian Alt, and Florian Mathis. Bans: Evaluation of bystander awareness notification systems for productivity in vr. In *Network and Distributed Systems Security (NDSS) Symposium*, 2023.

[36] Shrirang Mare, Franziska Roesner, and Tadayoshi Kohno. Smart devices in airbnbs: Considering privacy and security for both guests and hosts. *Proc. Priv. Enhancing Technol.*, 2020(2):436–458, 2020.

[37] Karola Marky, Nina Gerber, Michelle Gabriela Pelzer, Mohamed Khamis, and Max Mühlhäuser. "you offer privacy like you offer tea": Investigating mechanisms for improving guest privacy in iot-equipped households. 2022.

[38] Karola Marky, Alexandra Voit, Alina Stöver, Kai Kunze, Svenja Schröder, and Max Mühlhäuser. " i don't know how to protect myself": Understanding privacy perceptions resulting from the presence of bystanders in smart environments. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*, pages 1–11, 2020.

[39] Mark McGill, Daniel Boland, Roderick Murray-Smith, and Stephen Brewster. A dose of reality: Overcoming usability challenges in vr head-mounted displays. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 2143–2152, 2015.

[40] Francesca Mosca and Jose M Such. Elvira: An explainable agent for value and utility-driven multiuser privacy. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, pages 916–924, 2021.

[41] Josh Nadeau. Banking and Finance Data Breaches: Costs, Risks and More To Know. https://securityintelligence.com/articles/banking-finance-data-breach-costs-risks/, 2021.

[42] Kavous Salehzadeh Niksirat, Evanne Anthoine-Milhomme, Samuel Randin, Kévin Huguenin, and Mauro Cherubini. "i thought you were okay": Participatory design with young adults to fight multiparty privacy conflicts in online social networks. In *Designing Interactive Systems Conference (DIS)*, 2021.

[43] Jan Nolin and Nasrine Olson. The internet of things and convenience. *Internet Research*, 26(2):360–376, 2016.

[44] Farzad Nourmohammadzadeh Motlagh, Seyed Ali Alhosseini, Feng Cheng, and Christoph Meinel. An approach to multi-party privacy conflict resolution for co-owned images on content sharing platforms. In *Proceedings of the 2023 8th International Conference on Machine Learning Technologies*, pages 264–268, 2023.

[45] Joseph O'Hagan, Mohamed Khamis, Mark McGill, and Julie R Williamson. Exploring attitudes towards increasing user awareness of reality from within virtual reality. In *ACM International Conference on Interactive Media Experiences*, pages 151–160, 2022.

[46] Joseph O'Hagan, Pejman Saeghe, Jan Gugenheimer, Daniel Medeiros, Karola Marky, Mohamed Khamis, and Mark McGill. Privacy-enhancing technology and everyday augmented reality: Understanding bystanders' varying needs for awareness and consent. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(4):1–35, 2023.

[47] Joseph O'Hagan and Julie R Williamson. Reality aware vr headsets. In *Proceedings of the 9th ACM international symposium on pervasive displays*, pages 9–17, 2020.

[48] Seonghun Park, Jisoo Ha, Jimin Park, Kyeonggu Lee, and Chang-Hwan Im. Brain-controlled, ar-based home automation system using ssvep-based brain-computer interface and eog-based eye tracker: A feasibility study for the elderly end user. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:544–553, 2022.

[49] Sunyup Park, Anna Lenhart, Michael Zimmer, and Jessica Vitak. " nobody's happy": Design insights from {Privacy-Conscious} smart home power users on enhancing data transparency, visibility, and control. In *Nineteenth Symposium on Usable Privacy and Security (SOUPS 2023)*, 2023.

[50] Alfredo J Perez, Sherali Zeadally, Scott Griffith, Luis Y Matos Garcia, and Jaouad A Mouloud. A user study of a wearable system to enhance bystanders' facial privacy. *IoT*, 1(2):13, 2020.

[51] Sandra Petronio. Brief status report on communication privacy management theory. *Journal of Family Communication*, 13(1):6–14, 2013.

[52] James Pierce, Claire Weizenegger, Parag Nandi, Isha Agarwal, Gwenna Gram, Jade Hurrle, Hannah Liao, Betty Lo, Aaron Park, Aivy Phan, et al. Addressing adjacent actor privacy: Designing for bystanders, co-users, and surveilled subjects of smart home cameras. In *Designing Interactive Systems Conference*, pages 26–40, 2022.

[53] Blaine A Price, Avelie Stuart, Gul Calikli, Ciaran Mccormick, Vikram Mehta, Luke Hutton, Arosha K Bandara, Mark Levine, and Bashar Nuseibeh. Logging you, logging me: A replicable study of privacy and sharing behaviour in groups of visual lifeloggers. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(2):1–18, 2017.

[54] Leslie Ramos Salazar. Be Careful What you Post: Social Media and Reputation, 2021. https://profspeak.com/be-careful-what-you-post-social-media/, 2021.

[55] Kavous Salehzadeh Niksirat, Diana Korka, Hamza Harkous, Kévin Huguenin, and Mauro Cherubini. On the potential of mediation chatbots for mitigating multiparty privacy conflicts-a wizard-of-oz study. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–33, 2023.

[56] Florian Schaub, Rebecca Balebako, Adam L Durity, and Lorrie Faith Cranor. A design space for effective privacy notices. In *Eleventh symposium on usable privacy and security (SOUPS 2015)*, pages 1–17, 2015.

[57] Katrin Scheibe, Franziska Zimmer, Kaja Fietkiewicz, and Wolfgang Stock. Interpersonal relations and social actions on live streaming services. a systematic review on cyber-social relations. 2022.

[58] Samarth Singhal, Carman Neustaedter, Thecla Schiphorst, Anthony Tang, Abhisekh Patra, and Rui Pan. You are being watched: Bystanders' perspective on the use of camera devices in public spaces. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 3197–3203, 2016.

[59] John C Tang, Gina Venolia, and Kori M Inkpen. Meerkat and periscope: I stream, you stream, apps stream for live

streams. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 4770–4780, 2016.

[60] Parth Kirankumar Thakkar, Shijing He, Shiyu Xu, Danny Yuxing Huang, and Yaxing Yao. "it would probably turn into a social faux-pas": Users' and bystanders' preferences of privacy awareness mechanisms in smart homes. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2022.

[61] F Ted Tschang and Janusz Szczypula. Idea creation, constructivism and evolution as key characteristics in the videogame artifact design process. *European management journal*, 24(4):270–287, 2006.

[62] Froukje Sleeswijk Visser, Pieter Jan Stappers, Remko Van der Lugt, and Elizabeth BN Sanders. Contextmapping: experiences from practice. *CoDesign*, 1(2):119–149, 2005.

[63] Pamela Wisniewski, Heather Lipford, and David Wilson. Fighting for my space: Coping mechanisms for sns boundary regulation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 609–618, 2012.

[64] Yanlai Wu, Xinning Gui, Pamela J Wisniewski, and Yao Li. Do streamers care about bystanders' privacy? an examination of live streamers' considerations and strategies for bystanders' privacy management. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–29, 2023.

[65] Yanlai Wu, Yao Li, and Xinning Gui. " i am concerned, but...": Streamers' privacy concerns and strategies in live streaming information disclosure. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–31, 2022.

[66] Kaihe Xu, Yuanxiong Guo, Linke Guo, Yuguang Fang, and Xiaolin Li. My privacy my decision: Control of photo sharing on online social networks. *IEEE Transactions on Dependable and Secure Computing*, 14(2):199–210, 2015.

[67] Yaxing Yao, Huichuan Xia, Yun Huang, and Yang Wang. Free to Fly in Public Spaces: Drone Controllers' Privacy Perceptions and Practices. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 6789–6793, New York, NY, USA, 2017. Association for Computing Machinery.

[68] Tengfei Zheng, Tongqing Zhou, Qiang Liu, Kui Wu, and Zhiping Cai. Characterizing and detecting nonconsensual photo sharing on social networks. In *Pro-*

ceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, pages 3209–3222, 2022.

Appendix A   Examples of Participants' Original Designs.

(a)

(b)

(c)
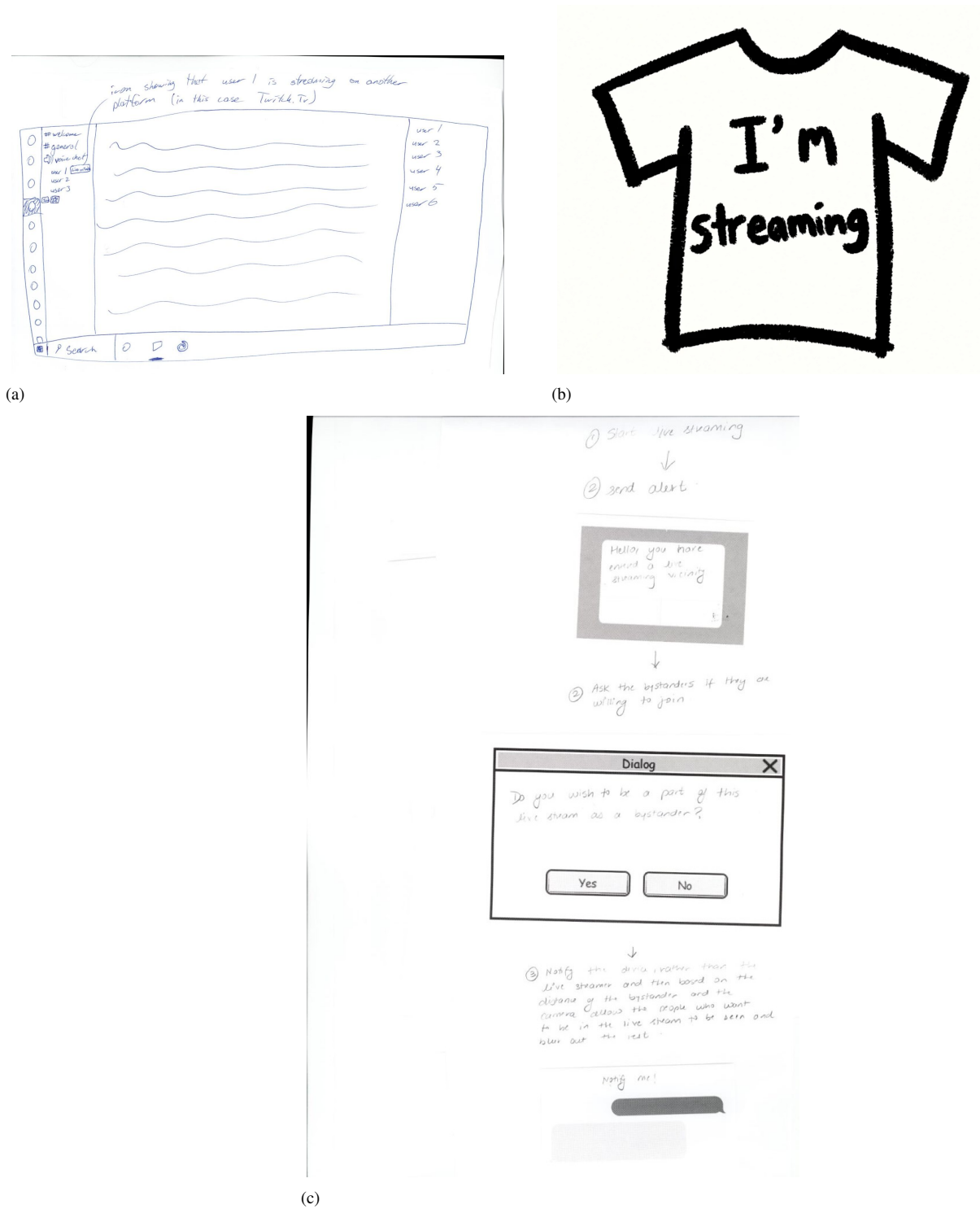
Figure 4: (a)provide gamers a link to the streamers' streaming channel (b)ask the streamer to wear t-shirts about the streaming (c)bystanders receive an automated SMS notification, and their consent responses are aggregated into a notification for the streamer.

Appendix B   Demographics of Participants.

Session	#	Gender	Occupation	Major	Bystander /Streamer	Streaming Topics	Stream Who	Where being Streamed	Streamed by Who
1(in person)	1	Male	Student	N/A	Bystander &Streamer	Valorent	Online Bystanders	In Game	Online Friend
	2	Male	Student	N/A	Bystander &Streamer	Outdoor Activities	Public Bystanders	At Bar	Unknown Streamer
2 (in person)	3	Male	Student	CS	Bystander	N/A	N/A	On Campus	Friend
	4	Male	Student	N/A	Bystander &Streamer	Overwatch	Online Bystanders	In Game	Online Friend & Opponent
3 (in person)	5	Female	Student	CS	Bystander	N/A	N/A	On Campus& In Farmers Market	Unknown Streamer
	6	Male	Student	CS	Bystander &Streamer	NBA 2K	Roommate	In Public	Unknown Streamer
4 (online)	7	Male	Student	CS	Bystander	N/A	N/A	In Game	Online Friend
	8	Male	Student	EC	Streamer	Casual Game	Parent	N/A	N/A
5 (in person)	10	Male	Student	EE	Bystander	N/A	N/A	In Game	Online Friend
6 (in person)	11	Male	Student	N/A	Bystander	N/A	N/A	On Campus	Unknown Streamer
7 (in person)	12	Male	Student	CS	Streamer	Rocket League	Online Friend & Roommate	N/A	N/A
8 (in person)	15	Female	Student	Psychology	Bystander	N/A	N/A	At Home	Roommate
	16	Female	Student	Psychology	Bystander	N/A	N/A	On Tennis Court	Sister
	17	Male	Student	CS	Bystander &Streamer	Teaching Coding	Family	At Friend's Home	Friend
	18	Male	Student	CS	Bystander &Streamer	Food	People in Restaurant	At Home	Roommate
	19	Female	Student	CS	Bystander	N/A	N/A	At Friend's Home	Friend
9 (in person)	21	Female	Student	Psychology	Bystander &Streamer	Teaching English	N/A	In Public	Friend
10 (in person)	22	Female	University Staff	Communication	Bystander &Streamer	Sport Game	Children	At Home	Children
11 (in person)	23	Male	Student	CS	Streamer	Sport Game	Public Bystanders	N/A	N/A
12 (online)	24	Male	Student	Game Design	Bystander &Streamer	Casual Game	N/A	At Home	Friend & Roommate
13 (online)	25	Female	Full-time Streamer	N/A	Bystander &Streamer	Singing &Dancing	Family	At Restaurant	Unknown Streamer

"It was honestly just gambling": Investigating the Experiences of Teenage Cryptocurrency Users on Reddit

Elijah Bouma-Sims
Carnegie Mellon University

Hiba Hassan
Carnegie Mellon University

Alexandra Nisenoff
Carnegie Mellon University

Lorrie Faith Cranor
Carnegie Mellon University

Nicolas Christin
Carnegie Mellon University

Abstract

Despite fears that minors may use unregulated cryptocurrency exchanges to gain access to risky investments, little is known about the experience of underage cryptocurrency users. To learn how teenagers access digital assets and the risks they encounter while using them, we conducted a multi-stage, inductive content analysis of 1,676 posts made to teenage communities on Reddit containing keywords related to cryptocurrency. We identified 1,409 (84.0%) posts that meaningfully discussed cryptocurrency, finding that teenagers most often use accounts in their parents' names to purchase cryptocurrencies, presumably to avoid age restrictions. Teenagers appear motivated to invest by the potential for relatively large, short-term profits, but some discussed a sense of entertainment, ideological motivation, or an interest in technology. We identified many of the same harms adult users of digital assets encountered, including investment loss, victimization by fraud, and loss of keys. We discuss the implications of our results in the context of the ongoing debates over cryptocurrency regulation.

1 Introduction

Over the last decade, cryptocurrencies, non-fungible tokens, and other crypto assets [28] have become very popular investments, especially among younger people. An online survey of 2,872 users conducted at the end of 2022 found that cryptocurrencies were the most popular type of investment among investors aged 18 to 25 [15]. Moreover, many survey respondents began investing at an extremely young age, with 25% reporting that they began investing as a minor. This

raises important security and privacy concerns: Private keys are difficult for even adult users to manage safely and any mistakes can result in immediate financial loss [47]. Most cryptocurrencies are pseudonymous, meaning that once a person's wallet address is known, many, if not all, of their past transactions can be identified [56]. Besides these usability and privacy issues, the cryptocurrency world has historically been rife with scams and other financial crimes that defraud users [7, 37, 59, 66].

Crypto asset investment also presents considerable financial risks. Many speculative crypto assets feature the characteristics of "gamblified" investments as defined by Newall et al. [63]: it is difficult for most to profit reliably [22, 26, 32, 80], it is attractive to users who are susceptible to gambling [43], and it presents the allure of out-sized profits. Even more established cryptocurrencies, like Bitcoin, are highly volatile, with the price of Bitcoin peaking around $65,000 in November 2021 before dropping to a three-year low of around $15,000 just a year later.[1]

There is good reason to hypothesize that minor teenagers (aged 13 to 17 years old) may be more vulnerable to these risks of cryptocurrency. Teenagers have developing brains, generally showing higher risk-taking and sensation-seeking behavior that declines with age [48, pp. 528–530]. They may also be affected more by certain motivational factors such as peer pressure [23] or influence from social media personalities [21], some of which have been implicated in illegally promoting crypto assets [7, 37]. Moreover, like other marginalized and vulnerable groups (e.g., immigrants [20], older adults [67], etc.), teens could be targeted by unique types of fraud tailored to their experiences.

The under-regulated nature of cryptocurrencies has also raised concerns [76] that teenagers may gain access to risky assets without parental oversight. While most exchanges require users to be 18 years of age or older to create an account,[2] it is

---

[1]Historical prices obtained from https://coinmarketcap.com/

[2]For example, the user agreement of the publicly traded cryptocurrency exchange, Coinbase, states that "To be eligible to use the Coinbase Services, you must be at least 18 years old." [19]

possible that teenagers could invest via unregulated exchanges with absent or lax Know Your Customer (KYC) policies.[3] Indeed, this method is advocated by online blogs [6, 49].

Despite these risks, little empirical evidence exists about how minor teenagers use crypto assets and the harms they encounter while investing. To address this gap, we analyzed posts about cryptocurrency from teenagers on Reddit, a US-based social media platform and link aggregation website. First, we used a set of 48 keywords to identify 4,979 posts likely to be about cryptocurrency in six Reddit communities ("subreddits") used by teenagers (e.g., /r/Teenagers, /r/ApplyingToCollege, etc.). We then performed inductive thematic analysis on a random subset of 1,676 posts, seeking to answer the following research questions:

1. **How do teenage users gain access to the crypto asset ecosystem?**
2. **What motivates teenagers to engage with crypto assets?**
3. **What types of harm do teenagers experience when using crypto assets?**

We ultimately identified 1,409 (84.0%) posts meaningfully discussing cryptocurrency. We found little evidence for the widespread use of unregulated exchanges, with minor teenagers most often claiming to use accounts in their parents' names to purchase cryptocurrencies. Some also received donations from other users or mined for cryptocurrency. Teenagers appear motivated to invest by profit, but some discussed a sense of entertainment, ideological motivation, or an interest in technology. Finally, we identified many of the same harms adult users of digital assets encountered, including investment loss, victimization by fraud, and loss of keys. Our main contribution is to provide the first insight into teenagers' interactions with cryptocurrency and the harms they encounter. We discuss our results in the context of ongoing debates about crypto asset regulation.

2 Background and Related Work

In this section, we review important background information and previous research related to our own, including 1) usable security and privacy research focused on crypto assets, 2) research about crypto assets and social media, and 3) children's online safety.

2.1 Cryptocurrency and Usable Privacy and Security

One of the main security and usability challenges of cryptocurrencies is the need to manage cryptographic keys. Rather than relying on a central intermediary to process transactions,

cryptocurrencies use public-key cryptography to verify transactions [60]. To prevent financial loss, users must ensure that they maintain access to their private keys or the underlying seed phrase while also ensuring that their (private) keys are protected from attackers. Cryptocurrency wallets are, therefore, highly vulnerable to accidental loss [72], social engineering attacks [89], or hacking [87]. Alternatively, users may rely on custodial services like wallet providers or exchanges that manage cryptocurrency on behalf of users. While custodial services lift much of the key management burden from users, they introduce new institutional risks (e.g., the service may be hacked as Bitfinex was in 2016 [68]).

Previous research has shown that cryptocurrency users can struggle to manage their cryptographic keys securely [1, 31, 47, 52, 88]. For example, Krombholz et al. [47] surveyed 990 Bitcoin users on their security and privacy practices, finding that nearly a quarter of users experienced some sort of financial loss due to user error (e.g., they formatted the hard drive containing their private key), system failure (e.g., dead hard drive, corrupted key file), or a malicious attacker (e.g., malware). Users' lack of conceptual understanding of how cryptocurrencies work may also contribute to these security lapses. Mai et al. [52] conducted an interview study with 29 participants (both current cryptocurrency users and non-cryptocurrency users) to learn about their mental models of cryptocurrencies in relation to security and privacy threats. They identified a number of misconceptions that could lead to financial loss, especially with respect to the function of private keys in cryptocurrency systems. For example, some participants did not realize that private keys were unique to each user and should not be shared with others. These misconceptions could be mitigated by improving the design of cryptocurrency wallets [25, 88] and abstracting away key management tasks [52].

Cryptocurrencies also present unique privacy risks. By necessity, all transactions are recorded on a public ledger, which can be viewed by anyone. While some cryptocurrencies (e.g., Monero, Zcash, etc.) attempt to offer greater privacy guarantees, the most popular cryptocurrencies, including Bitcoin and Ethereum, are merely pseudonymous. That is, there is no inherent link between a person's identity and the address that corresponds to their wallet, but anyone who knows that an address belongs to a particular person can find their transaction history on the blockchain. Moreover, users may be identifiable even if they take steps to hide their identity [5, 11, 56]. Users may overestimate the privacy guarantees of cryptocurrencies. Many participants in the study by Mai et al. [52] believed that transactions on the blockchain were encrypted and, therefore, could not be tracked. Similarly, almost a third of the participants in the study conducted by Krombholz et al. [47] incorrectly believed that Bitcoin was anonymous.

---

[3]KYC regulations vary by jurisdiction but generally require financial institutions to verify a user's identity to help prevent money laundering and other criminal transactions [30].

2.2 Cryptocurrency and Online Communities

Social media plays an important role in modern investing, including cryptocurrency. Social media platforms can help users to learn about new cryptocurrencies [41] and promote the adoption of their favorite projects [46,57]. Bad actors also use social media to manipulate the cryptocurrency market and promote scams [59,65,86]. For example, Nizzoli et al. [65] compiled a dataset of 50 million messages discussing cryptocurrency on Twitter, Discord, and Telegram, finding that Twitter bots were used to promote hundreds of different Telegram channels that facilitated Ponzi schemes and pump-and-dumps. Social media users with large followings, commonly known as "influencers" [40], have been implicated in illegally using their platforms to promote cryptocurrencies they own or were paid to promote without proper disclosure [7, 37], sometimes while actively investing against the very products they were promoting [45].

Reddit, in particular, has many communities focused on finance and investing (e.g., /r/WallStreetBets, /r/PersonalFinance, etc.), including many dedicated to discussing cryptocurrencies (e.g., /r/CryptoCurrency, /r/Bitcoin, etc.). Discussions of finance are not restricted to these communities, with cryptocurrency being the most popular topic across all of Reddit in 2021 [75]. A number of researchers have applied quantitative methods to study cryptocurrency discussion on Reddit, with a specific focus on the relationships with price movements [70, 71, 91, 95]. For example, Papadamou et al. [70] conducted an analysis of the relationship between Reddit activity and the price of cryptocurrencies, finding a strong cross-correlation between the number of posts mentioning a cryptocurrency and its price for 30 of the top-50 currencies by market cap. They also observed a correlation between average sentiment and price movement, with greater joy expressed during market upswings and anger expressed during market downturns.

Only a few researchers have applied qualitative methods to characterize the discussion of cryptocurrency on Reddit [17, 34, 44, 46]. Most relevantly, Childs [17] conducted a thematic analysis of the top 200 posts on /r/CryptoCurrency that mentioned the word "scam" in order to evaluate how users cope with fraud. He found that users attempt to prepare the community to deal with scams by providing resources (e.g., providing newcomer guides that discuss scams, sharing experiences of victimization, etc.) and establishing community norms to counter scammers (e.g., calling out suspected scam projects, identifying blockchain addresses associated with scams, etc.). He also found that users seem to promote the view that scams are an inevitable "cost of decentralisation." Johnson et al. [44] performed a thematic analysis of posts discussing psychological well-being, mental health, or gambling made to /r/CryptoCurrency during a decline in the cryptocurrency markets in 2022. They identified coping strategies users employ emotionally to handle the downturn, as well

as content that explicitly and implicitly connected cryptocurrency trading with gambling. To the best of our knowledge, no prior research has focused on online cryptocurrency discussions among teenagers.

2.3 Teenagers and Online Safety

Teenagers represent an important demographic to study due to their distinctive patterns of online behavior and susceptibility to digital risks. Modern American teenagers are "digital natives" [54] who have grown up in a world where digital devices and the internet are commonplace. The vast majority of American teenagers own or have access to a smartphone (95%) or a desktop/laptop computer (90%). Most also use some form of social media, with YouTube (93%), TikTok (63%), Snapchat (60%), and Instagram (59%) being the most popular. Only 14% of American teenagers use Reddit [4].

The teenage years are a critical psychological developmental stage. At that age, emotional and social networks in the brain mature faster than the prefrontal cognitive-control network, which continues to develop into the early to mid-20s. This renders teenagers particularly susceptible to heightened impulsivity, sensation seeking, and emotional reactivity. While there is a great deal of individual variance, adolescents often encounter challenges in performing executive function tasks requiring inhibition, planning, and future orientation. These developmental dynamics contribute to a general trend of increased risk-taking behaviors among teenagers that declines with age [48, pgs. 528 – 530].

Teenagers face many of the same security and privacy challenges as adults, such as poor password management [85], susceptibility to phishing [64], and difficulties managing both interpersonal and data privacy [50,55] online. Teenagers may experience these risks in ways distinct from adults. For example, Jia et al. [42] argue that some degree of privacy risk-taking (e.g., over-sharing information) on social media may help teenagers develop their privacy risk-coping strategies.

Teenagers also encounter some online risks more frequently than adults. For example, online sexual exploitation [3, 90] and cyberbullying [94] have been extensively studied in the scientific literature. No prior research has explored the types of safety risks that teens may encounter while using cryptocurrency.

3 Methods

We next describe the methods used to answer our research questions. We conducted an inductive thematic analysis of 1,676 posts containing keywords related to cryptocurrency from a set of Reddit communities used by English-speaking teenagers.

3.1 Dataset

Our study is based on posts and comments on Reddit. The platform is divided into thousands of user-created communities called "subreddits," which focus on particular topics (e.g., /r/politics for American politics) or identity groups (e.g., /r/gaybros for gay men). Users can submit posts that link to external websites or present original text and multimedia content. Users can vote to affect the visibility of posts, with an "upvote" boosting a post and a "downvote" lowering the post's visibility. Users can also comment on posts, with the order of comments determined by a similar voting scheme. Reddit is a common data source for academic research, particularly in computer science [73].

Rather than collect data directly from Reddit, we used an archived copy of the Pushshift dataset [8] with posts and comments from June 2005 to December 2022 (inclusive).[4] The dataset also includes the content of some posts that were deleted by users or removed by moderators, making it more complete than the data available directly from Reddit. Pushshift is widely used for academic research on Reddit [8,73].

Anyone 13 or older can register for Reddit, and an estimated 14% of American teenagers have used the platform [4]. For our content analysis, we focus on six communities that we assume are predominantly used by teenagers: /r/Teenagers, /r/ApplyingToCollege, /r/SAT, /r/ACT, /r/HighSchool, and /r/PSAT (henceforth referred to collectively as the "teenage subreddits"). /r/Teenagers describes itself as "the biggest community forum run by teenagers for teenagers," /r/ApplyingToCollege has over 1 million users and focuses on topics related to college admissions, with an emphasis on students enrolling directly out of high school. Similarly, /r/SAT, /r/ACT, and /r/PSAT are subreddits focused on discussion of college admissions exams that are most frequently taken by teenagers. Finally, /r/HighSchool is a subreddit that includes a wide range of content related to the secondary school experience. Previous studies [10, 16, 82, 93] have considered users' participation in Reddit communities like /r/Teenagers as an indication of a user being underage.

To identify posts that discussed cryptocurrency, we used a keyword-based sampling technique. Previous studies of Reddit have also used keyword-based sampling to study cryptocurrency [44, 70] as well as other topics [29, 92]. We first selected a heterogeneous set of keywords based on popular cryptocurrencies (e.g., Bitcoin, Ethereum, Dogecoin, etc.), cryptocurrency exchanges (e.g., Binance, FTX, etc.), interest-yielding services (e.g., Nexo, Blockfi, etc.), cryptocurrency gambling services (stake.com, Cloudbet, etc.), and concepts related to cryptocurrency (e.g., blockchain, decentralized, etc.). Our goal in selecting keywords was to identify a diverse set of

posts about cryptocurrency with a low false positive rate. The complete list of keywords and their frequencies can be found in Appendix B.

We then collected all posts that contained one or more of these keywords in either the title or the body of the post. We initially collected 6,408 posts; however, as discussed in the next subsection, several keywords were removed during codebook development. The final dataset contained 4,979 posts (0.06% of the total posts on the teen subreddits). For the thematic analysis, we selected a random sample of 1,676 (33.7%) to review.[5] We also collected all the comments ($n = 3,738$) associated with the posts in the sample. Most posts (80.0% of the sample) included only a single unique keyword. The most common keyword was some variation of "NFT," with 680 posts (40.6% of the sample) including at least one instance of this word.

3.2 Thematic Analysis

Thematic analysis was performed by two authors in multiple stages. To generate an initial codebook, the lead coder reviewed a subset of 100 posts from the 6,408 post dataset and performed inductive thematic analysis. The coder viewed the post title, post body, and all the comments associated with the post. If a post contained a link, the coder also considered this content if it was available.

Codes were selected to identify content relevant to our research questions and provide valuable categories for subsequent analysis. For example, we identified discussions of different types of behavior (trading, mining, using crypto for payment, etc.). We also identified different types of harm (e.g., monetary losses, crimes, and gambling). We also selected codes that characterized the kind of post (e.g., discussion of crypto-related news, joke/meme, question about crypto assets, etc.). Most posts had more than a single code assigned.

Some posts were recorded in the Pushift dataset with the body text replaced by "[removed]" or "[deleted]." This indicates that, before Pushift could collect the post, it was removed by moderators or deleted by the user who posted it. Often, it was still possible to assign a theme to these posts, although this analysis is necessarily less reliable. For example, a removed post titled "Artist Seed - The NFT project that births a metahuman through cryptoart" evidently promotes the Artist Seed project. Removed or deleted posts that could not be assigned a theme were coded as "Removed/Deleted (Ambiguous Content)." For example, a removed post with the title "Dogecoin" was given this code.

To help contextualize each post, we also viewed user and post flairs. Flairs are tags that can be added to a post or a user profile to indicate something about them. On /r/Teenagers, post flairs are required to indicate the type of content (e.g., "Meme," "Art," "Discussion," etc.).

---

[4]In response to changes in Reddit's policies in 2023, the Pushshift API is no longer publicly available to all users. We maintained an archive of the data available before this change and used it for this study.

[5]This sample size was selected based on the sample necessary to achieve a 99% confidence level with a 2.5% margin of error.

User flairs on `/r/Teenagers` are used to indicate a user's age. Flairs are unique to each subreddit. For example, on `/r/ApplyingToCollege`, user flairs are used to indicate a user's year in school (e.g., HS Freshman, College Senior, Graduate Student, etc.). When quoting users, we list their flair if available in Pushshift. All deleted posts and many removed posts had the author field replaced with the text "[deleted]" and the user flair removed. We are, therefore, unable to report age information for users with deleted posts.

During codebook development, some keywords were eliminated. For example, we realized that the keyword "ether" mostly resulted in false positives, as all posts returned by this keyword in the subsample were people misspelling the word "either." The keywords eliminated during this stage are listed in Appendix B with the note that they were removed.

After developing the initial codebook, the lead coder and another author collaboratively coded the remainder of the posts. The process was broken into blocks of 10% to 20% of the sample. For each block, the coders independently reviewed the sample and assigned codes based on the definitions in the codebook. The coders then met to compare their analyses, discuss differences, and select a consensus code for each post. Additional codes were added as needed throughout this process. During coding, the coders also recorded memorable or archetypal posts to serve as examples for the paper.

After all the posts were assigned an initial set of codes, the coders reviewed every post a second time to gain additional insights into the nuance of particular codes and refine the initial categorization. For example, for posts that were categorized as "Working on a crypto project," the coders enumerated the specific types of projects that users described. During this secondary analysis, the coders changed the codes assigned originally to some posts as needed. The final codebook with examples for each code can be found in Appendix A.

3.3 Ethical Considerations

Our research is not human subjects research, as we relied on publicly available data.[6] We recognize that social media users may not expect their posts to be reviewed and analyzed as part of research [27]. Moreover, teenagers, as members of a vulnerable population, deserve special consideration to ensure their protection. We do not report the names of users in the dataset. We will not publicly post our curated dataset, although interested researchers may contact the corresponding author to request an archived copy.

3.4 Limitations

The generalizability of our results is necessarily limited by the structure and demographics of Reddit. Due to the Reddit voting scheme, posts and comments that are more broadly

---

[6]Our study was submitted to our Institutional Review Board(s), which confirmed that human subjects research review was not necessary.
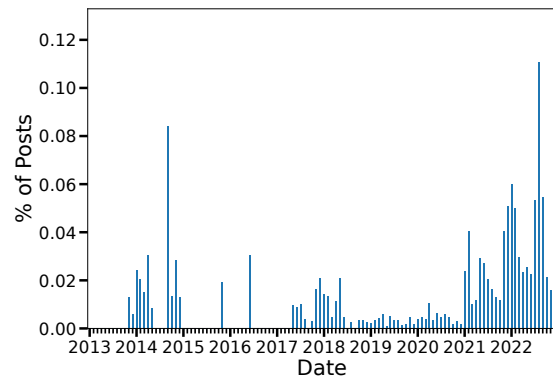
Figure 1: Frequency of posts in the random sample that contained relevant content each month, normalized by the number of posts in the teen subreddits per month. There were 8,803,440 posts on the teenage subreddits from January 1, 2013 to December 31, 2022. No posts in our sample were posted before 2013.

appreciated are more likely to be seen and interacted with. Reddit users are more likely male, with 20% of teen boys in the United States using Reddit as compared to only 10% of teen girls [4]. Moreover, we cannot verify the claims made by users, and some posts and comments may contain fabrications or exaggerations. The subreddits we investigated are not age-restricted, and some users are likely to be adults. Finally, we focused exclusively on English posts and users who do not speak English are, therefore, not represented.

4 Results

In this section, we describe the results from our qualitative analysis. We begin by giving an overview of the sample before discussing results relevant to each of our research questions.

4.1 Overview

Figure 1 shows the frequency of posts about cryptocurrency in our sample. To account for the varying number of posts per month, we normalized the number of posts each month by dividing by the total number of posts in the teen subreddits that month. The highest posting volume is associated with an increased discussion surrounding the Reddit NFTs, which were given away to many users in the second half of 2022 [53]. Only 267 (16.0%) of the posts in the random sample did not contain relevant content about cryptocurrency. Most of these posts (181) were false positives, which did not contain any mention of cryptocurrency. The rest of these posts (86) contained some mention of cryptocurrency, but in such a manner that they provided no useful information. For example, one user made a post in `/r/SAT` where they stated that they "want

Table 1: An overview of the results of coding the sample. The percentage for "# in Sample" refers to the proportion of the entire sample. The percentage for all other columns refers to the proportion of posts from each subreddit.

Subreddit	# in Sample (%)		# Irrelevant (%)		# Promotional (%)		# Removed/Deleted (%)	
/r/teenagers	1486	(88.7%)	204	(13.7%)	241	(16.2%)	497	(33.4%)
/r/ApplyingToCollege	109	(6.5%)	52	(47.8%)	10	(9.2%)	15	(13.8%)
/r/Sat	47	(2.8%)	2	(4.3%)	43	(91.5%)	42	(89.4%)
/r/ACT	15	(0.9%)	6	(40.0%)	7	(46.7%)	7	(46.7%)
/r/APStudents	11	(0.7%)	0	(0.0%)	3	(27.3%)	2	(18.2%)
/r/highschool	8	(0.5%)	3	(37.5%)	1	(12.5%)	2	(25.0%)
/r/psat	0	(0.0%)						
Total	1676	(100.0%)	267	(16.0%)	305	(18.2%)	565	(33.7%)

to apply to Computer engineering (preferring specialization in blockchain tech., cloud computing, data tech.)" While this indicates that they have some interest in crypto-related technology, it does not indicate anything about whether or not they currently use cryptocurrencies.

1,027 of the 1,409 relevant posts in our sample (72.9%) were made by users without user flair or with a flair that did not indicate their age (e.g., "2 MILLION ATTENDEE"). The most common age flairs identified the author as under 18 (385 or 81.7% of posts with age flair). Only 78 of the relevant posts (17.4% of posts with age flair) indicated that a user was 18 or 19.[7] The remaining 4 relevant posts (0.8% of posts with age flair) were made by users with the flair "OLD," indicating that the user was 20 years old or older. These age distributions should be viewed skeptically, as they are sparse and user-provided. Users may forget to change their flair after a birthday or purposefully lie about their age. Still, this result supports our assumption that most of the discussion in our sample comes from users under the age of 18.

Table 1 provides an overview of the posts in our sample, broken down by subreddit. The overwhelming majority of posts in the sample were posted on /r/teenagers (88.7%). This result is in line with the distribution of posts in the PushShift dataset, as /r/teenagers is by far the largest subreddit we examined. /r/PSAT contained no posts containing any of the cryptocurrency-related keywords. Across the entire sample, 33.7% of the posts were removed by moderators or were deleted by users prior to collection by PushShift. 67 of these posts (11.9%) had no code assigned (i.e., the probable content was ambiguous). The remaining 498 posts (88.1%) were assigned one or more codes, albeit at a lower confidence level than others. When reporting the frequency of codes in Appendix A, we specify the number that had removed or deleted content.

Table 2 shows the frequency of codes in our dataset. The most common post type in the sample were posts that pro-

moted a specific project or service (305 instances or 18.2%). These posts seemed fairly ineffective: 228 (74.8%) were removed or deleted, and most (200 or 66.6%) received no comments. Many of these posts advertised the same projects repeatedly, often using identical or near-identical text. The most commonly promoted projects were Pi Network (8.9% of promotional posts), Ethereum Name Service (7.8% of promotional posts), and Dogecoin (7.2% of promotional posts). Another common feature of these posts was some sort of give-away or sign-up bonus. For example, many posts promoting the Ethereum Name Service stated "Ethereum Name Service ($ENS) is Airdropping Tokens worth up to 5000$ for the first 1000 People To Claim it." Airdrops are free distributions of digital assets given away to promote a project [2]. When promotional posts were not removed right away, users often reacted negatively. For example, a post promoting Pi Network in /r/teenagers in August 2019 received negative replies from 5 different users, many calling it a scam (e.g., "That's 100% scam. Watch out").

As one might expect from a social media website, jokes and meme posts were also common. 254 posts (15.2%) were coded as being a joke or meme, mostly because they were explicitly labeled as such (i.e., using post flair). An additional 52 posts (3.1%) were tagged as probable sarcasm. We used this code when a post was not labeled as a joke but seemed so absurd as to be unbelievable or was otherwise clearly intended to be a joke. For example, one deleted post from 2019 to /r/teenagers stated, "$20 for a pic of my belly PM me accepting Bitcoin only (pic unrelated)." This post is unlikely to represent a genuine attempt at selling photos for Bitcoin. There was a great heterogeneity among the different types of jokes. However, a common theme seemed to be criticizing the intangibility of cryptocurrencies in general and NFTs in particular. For example, many users made jokes about the ability to "screenshot' or "right-click and save" NFTs. When we coded a post with "joke or meme" or "sarcasm," we refrained from applying any other codes to that post since we assumed the post does not reflect actual behavior.

Many of the relevant, non-joke posts were uninformative

---

[7]For this analysis, we assume that users with flair "HS Senior" are 18. We assume that users with flair indicating another class year (E.g., "HS Junior" are under 18.

for our research questions. For example, 69 posts (4.1%) discussed news related to crypto assets, such as the failure of the exchange FTX. While these posts demonstrate that users were interested in discussing crypto, they lacked information about users' experience with crypto assets.

Posts discussed a range of interactions with crypto assets, including obtaining/holding crypto assets (226 posts or 13.5%), mining cryptocurrencies (59 posts or 3.5%), selling or otherwise transferring crypto assets to others (30 posts or 1.8%), short term trading (25 posts or 1.5%), using cryptocurrencies for payment (25 posts or 1.5%), and gambling with crypto assets (13 posts or 0.8%). It is important to note that not all of these posts described actual, current behavior. For example, 10.2% (23) of the posts tagged as "discussion of obtaining/holding crypto" discussed a user's desire to obtain a crypto asset and 5.8% (13) conveyed a question about obtaining crypto assets.

Table 2: Most common codes assigned to posts in the dataset. Complete definitions and examples can be found in appendix A

Code	Frequency (%)	
Explicit promotion of project	305	(18.2%)
Joke or meme	254	(15.2%)
Discussion of obtaining/holding crypto	226	(13.5%)
Irrelevant	181	(10.8%)
Reddit NFT	156	(9.3%)
Giveaway	145	(8.7%)
Criticism of crypto	110	(6.6%)
Subjective question about crypto	105	(6.3%)
Irrelevant, Crypto mentioned in passing	86	(5.1%)
Discussion of news in crypto	69	(4.1%)

4.2 RQ1: How do teenage users access crypto assets?

The most common way for teenagers in our sample to gain access to cryptocurrency services was apparently by creating accounts in the name of their parents or another trusted adult. While most posts about obtaining crypto assets did not explicitly address how the user accessed the crypto ecosystem, we observed users discussing this when they asked others how they could access crypto assets. For example, in November 2021, one user posted a thread to /r/teenagers titled "Are you able to get into the stock market and cryptocurrency as a teenager." The body of the post stated "I wanna buy one of those funny looking monkey pictures but I'm clueless on how it works." One 17-year old user replied "you can have your parents set up your account." A different 17-year old user stated "I trade crypto under my moms name but to my knowledge you have to be 18 at least where I live." In response to a deleted post titled "Had to sell a lot of my crypto,"

made to /r/teenagers in November 2021, one user wrote "Dude I am sorry... can you tell me why you had to sell it, how you got in, and where do you sell your crypto (I've been tring get in for a while but i dont know how)." A 15-year-old user replied, writing, "I'm not OP but for crypto I would use Binance. It's an online exchange with pretty low fees. You'd need your parents' permission though, since it requires you to be over 18 and provide a drivers license..." Similarly, in reply to a post promoting Bitcoin in February 2021, one user commented "I wanna use Blockfi and earn interest but I'm not 18 :(((((((." A 15-year user replied with advice, stating "If you have a parent, you could do it with their account." Purchasing cryptocurrency through a parent's account is likely the safest way for teenagers to access the crypto markets, assuming they supervise their child's investments.

Users also discussed mining cryptocurrency to use it for speculation. For example, one 17-year-old user made a post to /r/teenagers in March 2021 titled "Decided to start mining for eth and made some btc out of it. Time to hold and watch it raise!" The image attached to the post showed $100 worth of Bitcoin in a wallet. Mining refers to the process of generating cryptocurrency by running a computer that validates transactions on a blockchain. Once mined, cryptocurrency could be transferred to an exchange and traded for other cryptocurrencies. This is presumably what the user meant when they said they were "mining for eth and made some btc." If an underage user were to use a decentralized exchange or other service that does not deal in US dollars or other fiat currencies, they could avoid age restrictions and KYC regulations. Another user more explicitly described this process in a post to /r/ApplyToCollege in July 2017: "Recently I started to get engaged with the cryptocurrency market. Ive been building multiple mining rigs and buying/sellimg the currencies in an attempt to make money. Would this count as an extracurricular?" Throughout our analysis, we encountered many posts similar to this, where users described some kind of involvement in crypto to see if it was worth discussing in their college applications.

Mining was also recommended to some users who inquired about obtaining cryptocurrency as a minor. For example, a 17-year-old user posted a thread to /r/teenagers titled "Is there anyway to buy crypto without being 18." The body of the thread stated "Since i cant participate in stonks without being 18 crypto i legally can, is there any service that doesnt ask for age verification?" Along with several replies recommending that the user get help from his parents or another adult, one 16-year-old user replied that they could "GPU mine for a while... and trade with that."

Some users also gained cryptocurrency via gifts from other users on Reddit. For example, in May 2021, a 15-year-old user posted to /r/teenagers "thank you to whoever tipped me dogecoin like a few months back." The body of the post added "i managed to reinvest in crypto and now have £290 from like £40" We also found a few examples of threads where users

gave away small amounts of cryptocurrency to users who commented in /r/teenagers. In January, 2014 a 16-year-old made a post titled "[Other]I'm back with another bitcoin giveaway!" The poster used the /u/changetip[8] bot to gift users fractional amounts of bitcoin. The post received over 900 comments. Similarly, in September 2014, a user made a post to /r/teenagers with the title "Free Bitcoin - Just Comment." The thread also received hundreds of comments.

This kind of post was rare. Most of the threads tagged with the "giveaway" code were promotional posts advertising some sort of airdrop or sign-up bonus given out by a service. All of the peer-to-peer giveaways occurred in 2014. Users directly donating to others could be motivated by a sense of philanthropy; however, donating small amounts of cryptocurrency directly to other users is also a good way to introduce new users to the ecosystem. Indeed, the poster in the January 2014 giveaway explicitly stated this, writing, "...I am doing this giveaway to get my peers involved in bitcoin. I believe in bitcoin as a currency, and not just an investment, and I think it could benefit other teens..." As Bitcoin became more mainstream, it became less necessary to introduce users through free giveaways.

Some teenagers created their own crypto projects. For example, one user made a post to /r/ApplyingToCollege in August 2018 where they asked if they should update colleges about a cryptocurrency project they launched: "So I just recently finished making my own crypit currency with a relatively high market cap of 100k ish. I didn't mention this on my app at all (besides interest in crypto) because I didn't think I would get it done in time, but I did. Would it be worth emailing my colleges about this new stage in my life..." The open-source nature of major cryptocurrencies makes it relatively trivial to fork an existing project, tweak some properties, and give the currency a new name. Dogecoin, for example, began as a fork of an existing project (LuckyCoin) [18].

We found little evidence for teenagers using unregulated exchanges to purchase cryptocurrency. We only found one comment or post where a user stated that they used an unregulated exchange to avoid "know-your-customer" (KYC) practices. In reply to a comment expressing frustration about being unable to purchase cryptocurrency, one user in mid 2021 stated "Lol I'm a minor and I bought Bitcoin using a non kyc site, and a vpn..." Using a VPN could allow users to avoid KYC requirements and other regulations that are only enforced for users who appear to be in a particular jurisdiction. Some exchanges, such as Binance International, were complicit in helping US-based users avoid geo-blocking by recommending the use of a VPN [51]. We did find another comment where a user recommended Binance International in a discussion where a different user shared that they were interested in getting started in cryptocurrency: "yeah go for it man, hmu if U need help, im using binance international."

This user did not indicate that they were under 18 nor did they state that they were using Binance International to avoid regulations.

There may be more users who used unregulated exchanges but did not discuss it for various reasons (e.g., they may not want to admit to trying to circumvent exchange rules). Our keyword-based sampling approach may have also prevented us from finding more examples. Still, finding only a single example of the use of an unregulated crypto exchange suggests that the practice is less common than the other ways minors may obtain crypto assets.

4.3 RQ2: What motivates teenagers to engage with crypto assets?

A desire to realize relatively large, short-term profits seemed to be the most common motive for teens acquiring crypto assets. The most valuable insight into users' motives came from 23 of the posts coded with "discussion of obtaining/holding crypto" where users discussed their intention or desire to purchase crypto assets. The potential for large, relatively short-term gains was commonly brought up. For example, one 14-year-old user made a post to /r/teenagers in March 2021, stating, "Hey guys so my parents are giving me 100 dollars to invest in whatever I want... I think a good crypto to buy is Bitcoin... I've been tracking it... it is fairly stable and usually doesn't go down a lot and since by 2030 my investment of 100 will grow ten fold if I buy it now." A different 14-year-old user made a post to /r/teenagers in June 2021, writing "I'm becoming obsessed. I wanna transfer some of my personal money into Bitcoin so I can have enough for a car when I'm 17 in 3 years. Anyone else involved at all in crypto?" In reply, several users recommended against trading crypto. However, a 15-year-old user shared an anecdote that might reinforce the sense of easy profits: "I put $500 on doge when it was 0.04 and when it hit 0.72 it became $5,000."

Users actively engaged in crypto asset investment discussed how profitable it seemed to be compared to other ways to make money. For example, in August 2021, a user made a deleted post to /r/teenagers titled "Lmao. Trading stocks seems like a scam when I see shit like this. With Crypto trading I've seen 50-100% profits in literally minutes. Still learning but starting with small amounts. Like 18/20 of the trades I made so far were at least 50% profit." The post linked to a picture of text explaining that day trading stocks are typically unprofitable. Similarly, in November 2021, a user made a post to /r/teenagers titled "If Shiba Inu lists on Robinhood I'll have enough money to move out of my parents house." The body of the post stated "And yes that's my only hope... because dog crypto coins have made me more profit in a week than minimum wage pays in 8 months."

Short-term profit potential was also frequently emphasized in posts promoting crypto asset investing. For example, one user wrote a post to /r/teenagers in January 2021, stating

---

[8]This bot is now defunct. While operating, it allowed users to gift cryptocurrency using commands in Reddit comments [39]

"My advice to everyone here is to invest, whether it be in cryptocurrency or the stock market... Cryptocurrency (like bitcoin) is more wild and fluctuating than the stock market, so it's riskier... If you do it right, a measly $1K can pay your entire college fund in a few years." Similarly, an 18-year-old user wrote in a post to /r/teenagers in May 2021, "...I have reason to believe that Safemoon is going places... this crypto could rocket up, and if you invest now you could make a good profit if it does get anywhere..." This example is particularly disturbing, as the company and founders of SafeMoon were subsequently charged with defrauding investors [78].

Interestingly, several users described being pushed by their parents to develop NFTs based on a desire for outsized profits. For example, a 14-year-old user made a post to /r/teenagers in July 2022 titled "Any Crypto Teens here to answer this?" The body of the post shared "I do art as a hobby, and my dad's been telling me to turn them into NFTS. I've refused twice, but on the third time, he told me that I'm missing out on the chance to make him and I millionaires. Do I go for it or not?" Similarly, a 14-year-old user made a post to /r/teenagers titled "So uh." in January 2022. The body of the post stated "My parents are asking me to make NFTs and sell them because 'you have some knowledge about it, why not go and make some money like him...'" The post linked to an article about a computer science student who made millions of dollars from selling an NFT collection [24].

Users and their parents' desire to profit off crypto assets is generally misguided. While some popular NFT projects are very profitable, most make little to no money [69]. The story is similar in cryptocurrency markets. Most retail investors sold at a loss following the collapse in cryptocurrency prices in 2022 [22, 26], and many fall prey to institutional investors [80]. Short-term profits may trick users into feeling that they are skillful or have discovered a successful strategy, however, the ambiguity of the crypto asset market makes it difficult for traders to consistently profit over time [32]. Crypto asset trading may appeal particularly to teens due to their limited wealth and low incomes. A modest gain from buying and holding traditional equities may seem trivial to a person who only has a few hundred dollars to invest.

Some users discussed crypto investing as a form of obsessive entertainment akin to gambling. For example, one user made a post to /r/teenagers in November 2021 titled "I invested like $100 into crypto and [can't] stop looking at the charts and stuff." In response to a commenter who stated that "The market is too volatile imo," the original poster wrote "That makes it fun." A 17-year-old user replied to the thread, sharing the same experience: "I got interested in it a year ago and I couldn't stop looking at charts even during online classes for months." Another user shared that they went on a "...wild rally of not sleeping and keeping my eye on the charts." Increased checking of markets has previously been associated with investment in crypto assets [35].

User justifications for purchasing high-risk assets also mentioned the "fun" of the activity alongside the potential for a high reward. For example, in July 2021, a user made a post to /r/teenagers titled "i invested $1000 into a dog meme crypto haha." In the body, they explain "...I saw a lot of videos about it on tiktok and want to make some money so decided to stick $1,000 into a cryptocurrency meme coin called hoge lol... whether I become a millionare or make nothing it would be a fun experiment." In the comments, they stated they were under 18. Similarly, in response to a post promoting a project called PhunWallet in January 2021, one 17-year-old user wrote, "im interested, just for fun yk... I'm more into well established cryptos like btc,eth,sol,doge,aval,xrp,xlm,etc etc" The original poster replied stating, "Yeah this is a start up, costs nothing to get into this... maybe in 5 years it could be worth something."

The comparison with gambling is not just inferred. As Johnson et al. [44] observed, users in our sample directly compared the experience of crypto trading to gambling. For example, an 18-year-old user wrote in a post to /r/ApplyingToCollege in "...I got into investing around that time as well, and I flipped $10 to around $2.7k after multiple trades. It was honestly just gambling..." Similarly, a 17-year-old user responding to a critique of the riskiness of crypto assets wrote "u see risk = profits... cryptos are like gambling rn also its the best way to pay and receive money without paying extra because of tax... anyways if it gets me money i'mma do it."

The design of real-time trading services may contribute to the entertainment appeal of crypto asset investing. The continuous feedback and speed of transactions facilitated by these services can make them more engaging. Additionally, so-called "gamification" techniques (e.g., investing leaderboards, rewards points, etc.) are commonly used in trading services [74]. These have been shown to increase the frequency of trading, potentially leading to harm [12].

We also observed motives for engaging with crypto assets besides profit. A few users shared ideological justifications for using cryptocurrency, especially in posts promoting cryptocurrency use in general. 35 (2.1%) posts were coded as "General promotion of crypto." For example, one user made a post to /r/teenagers in February 2022 titled "Dudes fr never gave a single fuck about the environment until they want to use that as an excuse to hate crypto." The body of the post read "It's so ignorant too, banks use more energy and cause a lot more pollution than crypto. We're trying to phase out evil and corrupt banks, we can't do that without using any fucking power..." This kind of justification for the use of cryptocurrency has its roots in the first posts from the anonymous creator of Bitcoin who expressed a deep distrust of the conventional financial system [62].

A few posts, particularly in /r/ApplyingToCollege, discussed crypto assets in relation to a more general interest in technology. For example, one user wrote in a post in July 2018, "I kind of have a passion for home automation/raspberry pi and arduino projects. Using these mini computers (running

python and c++) I've made my own automatic door lock, a weather sensor and display, ... and a bot that manages and displays crypto prices." Another user made a post in October 2022, describing how they were the co-founder of their school's "Intel-Tech Club" in 11th grade. They explained that, as a leader in the club, they "...hosted Fintech, Business sessions; conducted bimonthly contests; hosted seminars on AI, Blockchain, ML, AR/VR, etc." Ultimately, this kind of purely technical interest in crypto assets was rarely discussed.

The Reddit collectible avatar NFTs introduced many users to crypto assets. The Reddit NFT was by far the most discussed crypto asset in our sample. 156 posts (9.3%) were coded as being about the Reddit NFT. Moreover, 99 (43.8%) of the posts coded as "discussion of obtaining/holding crypto" were about the Reddit NFT. Bitcoin (26 posts or 11.5%) and Dogecoin (17 posts or 7.5%) were the next most popular crypto assets that users discussed holding.

Reddit NFTs are hosted on the Polygon blockchain, and allow holders to uniquely customize their Reddit avatar. The NFT was given out for free to some on Reddit [53] in late 2022. In the wake of this giveaway, many users discussed receiving the avatar on /r/teenagers, with dozens of posts featuring images of the avatars titled things like "I got one of those free NFT avatars." and "I got random free NFT avatar while surfing the home page :/ and it is weird af." Some users reacted negatively, particularly to the idea of the avatars being NFTs. For example, one user made a post in August 2022, stating "I just got a fucking reddit nft avatar and I feel like a shitty crypto bro."

Unlike other crypto assets, users interested in purchasing a Reddit NFT discussed their aesthetic value. For example, a 15-year-old user wrote a post titled "am i the only one who's tempted to buy a reddit nft." By way of explanation, they commented "GUYS THEY LOOK COOL I DONT HAVE REDDIT PREMIUM I CANT MAKE MY [AVATAR] COOL LIKE URS." Similarly, a 13-year-old user made a deleted post to /r/teenagers in September 2022, writing "I want Reddit nft." In response to another user they explained "I want the reddit avatar items that come with the nft." These posts suggest that the Reddit NFT may be understood more like a cosmetic item in a video game rather than a financial asset.

4.4 RQ3: What types of harm do teenagers experience when using crypto assets?

Users discussed a variety of harms, including fraud victimization, wallet loss, and financial losses from poor investments. 27 posts (1.6%) were coded as including "crime," including crimes perpetrated by users and crimes victimizing users. Some crimes were only tangentially related to cryptocurrency. For example, several users discussed accounts being compromised to post spam related to crypto assets. This type of anecdote appears in our dataset because of the inclusion of crypto-keywords, however, this example of abuse is not related to the use of crypto assets.

The most common type of crime that victimized users was extortion facilitated by cryptocurrency. For example, one teen made a post titled "What do I do?" to /r/teenagers in 2020 asking for advice on a "sextortion" message he received: "I just got an email about something very weird. Someone emailed me some with the name of my password... it said that if I did [not] pay this certain address $2000 dollars in bitcoin it would send 3 random people in my contact a video of me wanking. Idk if this is real or just a scam to get money out of me. I am only 14 and don't know what to do. Please help me..." Thankfully, this user's post received several comments reassuring them that the extortion attempt could be safely ignored. For example, one 14-year-old user wrote "Lol it's a scam. Everyone and their mother has gotten that e-mail. You should change your password tho because it was leaked..." Bitcoin and other cryptocurrencies are often used to facilitate extortion and other scams [66], as they are pseudonymous and lack payment reversal mechanisms like credit card chargebacks [61].

Technical aspects of cryptocurrency were rarely discussed; however, we observed issues with key management. The giveaway posts (see subsection 4.2) featured the most discussion of wallets and key management, as users needed to be on-boarded into the ecosystem to receive their gift. Most of this discussion was vague and non-specific, providing little insight into users' key management practices.

Key loss was mentioned in a small number of posts and comments in the dataset. For example, a 15-year-old wrote in a post to /r/teenagers in January 2021 "I had $0.60 or whatever in ETH a long time ago, now it's worth $18 and I can't find the private key :(But i did find the key to an account with $11 of bitcoin so that's cool." This comment is typical of what teenage users described when discussing key loss: small amounts of forgotten money that became valuable in retrospect. More often than dedicated posts about key loss, users mentioned losing wallets in unrelated contexts. For example, a 16-year-old wrote in a giveaway thread, "About a year or so ago, I decided to try bitcoin mining, but my computer was pretty lame... I mined like one whole dollar! I wish I still had that hard drive :'(" This type of user error leading to financial harm is similar to that discussed by Krombholz et al. [47].

Users also shared experiences with investment and trading losses. More posts were coded as discussing profit (51 or 3.0%) than loss (17 or 1.0%). This likely does not reflect the overall frequency of investment losses relative to profits, as users may be too ashamed to discuss their losses. A few users shared extreme examples of loss. For example, a post titled "Screenshot of my crypto portfolio in October vs what it is now" in /r/teenagers in July of 2022 showed images of $18,000 in value dropping to about $100. Only the value of the assets was shown in the screenshots, so it is difficult to tell what type of investments this user made or if they engaged in active trading. A 17-year-old user shared an experience of

making a huge gain on paper, before losing most of it. The post was made to /r/teenagers in April 2022 with the title "I made $160K on crypto and my social anxiety dissapeared. Then I lost it and It's back but worse lmao." They explained in the body of the post that "...i made a high risk bet that paid off... I was worth 160k for like a couple weeks and i felt like the fucking man, my social anxiety literally crumbled i was so confident..." In the comments, they explained that they started with a few hundred dollars and provided a screenshot of their portfolio, which consisted of a handful of obscure cryptocurrencies.

Most losses from investment were much smaller than shown in these examples. For example, on /r/teenagers in December 2021, one 16-year-old user wrote a post titled "I bought yesterday shib for 50$ and it dipped in the night." The body of the post added "Pain. And btc also dipped from 170 I had mined to 140. Ehh just remember HODL."[9] In response, a 14-year-old user shared "my portfolio literally went from $9.5k to $7.5k this week lol." These losses are the inevitable flip side of engaging in high-risk investing for easy, short-term profits, particularly in hard-to-predict markets like those for crypto assets. While some users will be lucky enough to pick the right coin and sell it at the right time, others will experience losses. While many of the losses we observed are small in absolute terms, they may represent a large proportion of a teenager's total wealth.

We found several examples of teenagers losing money to pump-and-dump schemes. A pump-and-dump is where an investor or group of investors drives up the price of an asset through false or misleading statements, hype, or other manipulative tactics (the "pump"). Once the price has been inflated to a certain level, the perpetrators sell their holdings (the "dump"), causing the price to plummet and leaving many investors with losses. Pump and dump schemes are illegal in most jurisdictions. In a thread on /r/teenagers in April 2014, a 16-year-old user commented. "Want to know the struggle? I lost 47k on paper with cryptocurrencies. Yes, USD." In response to a reply, they elaborated "So, when I was dumb and gullible in the world of crypto earlier this year, I would try to ride these pump and dumps that people on twitter/IRC would do..." They explained the concept of a pump and dump before concluding with the statement "I was a bag holder in a twitter pump and dump of BlackCoin since early February. I had 180k of them. BlackCoin reached between 0.0008 and 0.009 at one point and currently hovers around 0.0005. If I had sold at 0.0008, I would have had 144 btc, a 7200% ROI. Actually looking at it now, that'd be about 72k USD. Instead I sold at 0.00008, taking a near 60% loss." Their comment suggests that they participated in multiple pump-and-dump schemes before the loss they discussed.

Another thread posted to /r/teenagers in February 2021

[9]The word "HODL" (a joking misspelling of the word hold) alongside the phrase "diamond hands" are commonly used in crypto communities to encourage users to hold onto assets, regardless of price activity [33].

also describes a pump-and-dump, although the user seems to understand less about what occurred. In the body of the post, they explain that they participated in a pump and dump: "Today a thing called pump(a lot of people buy a coin at the same time so the price go up and the you sell it have like 200% of profit) was made and i said ok this Is my moment... When the name of the coin was released I run to my investing app and spend 20€ in that coin... I wait to the price to raised and then I sell it. I don't no how but I finished losing money." This user seems to have participated in a pump and dump group similar to those described in Nizzoli et al. [65] but without the awareness that they risked losing their investment. These examples highlight the risk of uneducated and inexperienced teenagers participating in under-regulated markets.

One user claimed to have taken a leadership role in market manipulation. Using a one-time-use or "throwaway" account, they posted a thread to /r/ApplyingToCollege titled "How should I describe a crypto extracurricular?" requesting advice on how to discuss running a pump-and-dump group on college applications. The body of the post stated "I run a discord server... that has a lot of members and can move prices on obscure crypto (think <15 million market cap). I started with a small sum of ∼$40,000 from previous investments and small gifts, and have made nearly 100 times profit. I spend nearly all of my spare time doing this, so my grade in my foreign language dropped (from A to B) and I don't have any other ECs to put on my application other than really trivial things." Teenagers have previously been identified as perpetrators in various types of cybercrimes, including high-profile security breaches (e.g., "MafiaBoy" [38]). Leading a crypto pump-and-dump requires no significant resources or special skill, so while this description is concerning, it is unsurprising that a teen might engage in this kind of unethical and likely illegal behavior. The poster however insists their actions were legal, stating, "I cleared it with my parents and other family friends who are securities lawyers..."

Teenage users also described participating in cryptocurrency gambling. 13 posts (0.8%) were coded as containing content about gambling, with 4 of these posts promoting a service. These are distinct from posts where users described their trading behavior as analogous to gambling (see section 4.3), as they involve users discussing unregulated, online crypto casinos. These include centralized services and decentralized gambling applications hosted on a blockchain [14, 58]. In the US and Europe, the legal minimum gambling ages range from 18 to 25 [9], so minor teenagers are not legally permitted to gamble.

The most concerning example was a user who discussed what they describe as an addiction to both trading and gambling in a post to /r/ApplyingToCollege in November 2022. The post asked about how they should discuss cryptocurrency in their college applications. In the relevant part, they stated "...I did crypto/nfts/skins reselling during Quarantine. Most of my revenue came from altcoins mooning, such

as Ada and Solana... Probably made over \$250k+ in revenue (did this probably like 5-8 hours a day on average, I was fucking addicted)... Only came out with like \$3k in profit... I was a fucking gambling addict. I was addicted to Roobet, Bustabit, and Stake.com... I deeply regret doing crypto, it was such a mentally draining and useless thing...” Some users replied with similar experiences. For example, one high school senior wrote “i also had an addiction, crypto makes it way too easy to lose it all lol. glad to hear you're better...” Another user wrote “for reference, i made (roughly) 2 million on the crypto spike during 2020-2021 i am also a stake addict ~\$450k gambled i did also start a sports betting server and get involved with certain things i would rather not discuss, but overall i actually ended up profiting from gambling...”

Most examples of users discussing gambling were less extreme. For example, an 18-year-old user made a post in `/r/teenagers`, stating “Hey guys, I got into Dogecoin a while back and I've been able to gamble my way up for about \$5 from mining, to \$30... I've since lost all of that and now I'm sitting at about \$0.30 in doge, and I'm sad about it. I... will try to make the money back from the same gambling process with tips made on reddit...” It is startling that this user lost almost all their cryptocurrency gambling, but still felt that it was a reasonable way to make money. Incidents like this suggest that increased regulation in this space may be needed to protect teenagers and adults.

5 Discussion

This section discusses the implications of our results, particularly in contrast to adult populations. We also discuss ways to better protect users from the risks of crypto assets and the potential for future work investigating teen crypto users.

Underage teenagers on Reddit seem to most often gain access to crypto assets with the help of parents or other trusted adults rather than international services or decentralized exchanges, as was feared by Ryan [76] (RQ1). Teens' motivations for using crypto assets focus on profit, similar to adults (RQ2), as long-term investment, trading, or other forms of financial speculation are the most commonly reported uses of crypto assets in adult populations [81]. Parental pressure was a unique motivational factor we noted in several posts; however, this seems rooted in parents' sense that crypto could be highly profitable. Engaging in speculative investing and trading with crypto assets during a formative period may teach bad habits that could harm teens future financial success. We also identified many of the same harms experienced by adults [47], including financial loss due to poor key management, speculative investing, and even fraud (RQ3).

Underage teens' method of access to crypto assets seems to be the greatest factor that distinguishes them from adult users. While using an exchange account in the name of a parent may allow for some level of parental supervision, exchanges such as Coinbase and Binance do not support traditional “custodial accounts” [84] that are explicitly designed to enable parental oversight of finances. Implementing account types specifically designed for teens, perhaps with limited access to riskier assets, could promote safer crypto investment behavior.

Increased regulation of crypto assets may help protect both underage teenagers and adults in crypto markets. Some types of crypto assets may be regulated under existing laws governing securities and commodities. For example, in 2023, the US Securities and Exchange Commission (SEC) charged multiple crypto companies for operating unregistered securities exchanges, including Kraken [79] and Coinbase [77]. Forcing crypto assets to register as securities could increase transparency and allow greater market surveillance to prevent fraud. It would also promote legal clarity by unambiguously defining market manipulation, including pump-and-dump schemes, as illegal. However, treating crypto assets like traditional financial assets is far from a one-size-fits-all approach. As illustrated by the example of the Reddit NFT, not all crypto assets are purchased for profit-seeking motives, and some may be more akin to collectibles, which are generally not treated as securities or commodities. Moreover, some crypto assets that are purchased for their profit potential do not fit with the classic definition of securities [36].

Future work should explore teenage experiences with crypto assets more directly using human-subjects research. While we have confirmed that teens participate in the crypto markets, quantitative studies could determine the frequency of underage market participation and the various harms we identified. Qualitative studies could be employed to probe teens' motivations for investing in crypto and more deeply explore individual experiences. Such work is essential to further protect vulnerable populations like teenagers.

6 Conclusion

Our study sheds light on the largely unexplored area of underage cryptocurrency usage. Through an inductive content analysis of 1,676 Reddit posts from teenage communities, we found that teenagers predominantly utilize their parents' accounts to circumvent age restrictions and engage in crypto-asset transactions. Their primary motivation seems to be profit-seeking, although other motivations were discussed (e.g., ideological conviction). Our findings also highlight the risks inherent to this activity, including investment losses, fraud victimization, and cryptographic key loss. While somewhat limited by the unreliability of Reddit discussion, our research underscores the need for protective measures to safeguard young investors from potential harm.

References

[1] Svetlana Abramova, Artemij Voskobojnikov, Konstantin Beznosov, and Rainer Böhme. Bits under the mattress: Understanding different risk perceptions and security behaviors of crypto-asset users. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[2] airdrop.io, 2024. https://airdrops.io/, as of June 11, 2024.

[3] Sana Ali, Hiba Abou Haykal, and Enaam Youssef Mohammed Youssef. Child sexual abuse and the internet—a systematic review. *Human Arenas*, 6(2):404–421, 2023.

[4] Monica Anderson, Michelle Faverio, and Jeffrey Gottfried. Teens, Social Media and Technology 2023. Technical report, Pew Research Center, 2023. https://www.pewresearch.org/internet/2023/12/11/teens-social-media-and-technology-2023/.

[5] Elli Androulaki, Ghassan O Karame, Marc Roeschlin, Tobias Scherer, and Srdjan Capkun. Evaluating user privacy in bitcoin. In *Financial Cryptography and Data Security: 17th International Conference, FC 2013, Okinawa, Japan, April 1-5, 2013, Revised Selected Papers 17*, pages 34–51. Springer, 2013.

[6] Emma Avon. How to Buy Bitcoin Under 18? [Age Limit in 2023]. *CoinCodex*, July 2023. https://coincodex.com/article/30767/how-to-buy-bitcoin-under-18/, as of June 11, 2024.

[7] Timothy M Barry. # NotFinancialAdvice: Empowering the Federal Trade Commission to Regulate Cryptocurrency Social Media Influencers. *Ohio St. Bus. LJ*, 16:279, 2021.

[8] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839, 2020.

[9] Bet MGM Casino. The Legal Age for Casino Gambling Around the World, January 2023. https://casino.betmgm.com/en/blog/legal-age-casino-around-world/, as of June 11, 2024.

[10] Arpita Bhattacharya, Calvin Liang, Emily Y. Zeng, Kanishk Shukla, Miguel E. R. Wong, Sean A. Munson, and Julie A. Kientz. Engaging Teenagers in Asynchronous Online Groups to Design for Stress Management. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children*, IDC '19, page 26–37, New York, NY, USA, 2019. Association for Computing Machinery.

[11] Alex Biryukov and Sergei Tikhomirov. Deanonymization and Linkability of Cryptocurrency Transactions Based on Network Analysis. In *2019 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 172–184, 2019.

[12] Behavioural Insights Team (BIT). Digital Engagement Practices in Retail Investing: Gamifcation & Other Behavioural Techniques. Technical report, Ontario Securities Commission Investor Ofce, November 2022. https://www.osc.ca/sites/default/files/2022-11/inv_research_20221117_gamification-of-retail-investing_EN.pdf as of June 11, 2024.

[13] Antonio Briola, David Vidal-Tomás, Yuanrong Wang, and Tomaso Aste. Anatomy of a Stablecoin's failure: The Terra-Luna case. *Finance Research Letters*, 51:103358, 2023.

[14] Samuel Hoy VII Brown. Gambling on the Blockchain: How the Unlawful Internet Gambling Enforcement Act Has Opened the Door for Offshore Crypto Casinos. *Vand. J. Ent. & Tech. L.*, 24:535, 2021.

[15] FINRA Investor Education Foundation CFA Institute. Gen Z and Investing: Social Media, Crypto, FOMO, and Family, May 2023. https://rpc.cfainstitute.org/en/research/reports/2023/gen-z-investing, as of June 11, 2024.

[16] Robert Chew, Caroline Kery, Laura Baum, Thomas Bukowski, Annice Kim, Mario Navarro, et al. Predicting Age Groups of Reddit Users Based on Posting Behavior and Metadata: Classification Model Development and Validation. *JMIR Public Health and Surveillance*, 7(3):e25807, 2021.

[17] Andrew Childs. "I guess that's the price of decentralisation...": Understanding scam victimisation experiences in an online cryptocurrency community. *International Review of Victimology*, 2024.

[18] Usman W Chohan. A history of Dogecoin. *Discussion Series: Notes on the 21st Century*, 2021. `https://ssrn.com/abstract=3091219`.

[19] Coinbase. User agreement, November 2023. `https://www.coinbase.com/legal/user_agreement/united_states`, as of June 11, 2024.

[20] U.S. Federal Trade Commission. How To Avoid Immigration Scams and Get Real Help, 2023. `https://consumer.ftc.gov/articles/how-avoid-immigration-scams-and-get-real-help`, as of June 11, 2024.

[21] Nina Grgurić Čop and Barbara Culiberg. Business Is Business: The Difference in Perception of Influencer's Morality Between Generation Y and Z. In Francisco J. Martínez-López and Steven D'Alessandro, editors, *Advances in Digital Marketing and eCommerce*, pages 56–61, Cham, 2020. Springer International Publishing.

[22] Giulio Cornelli, Sebastian Doerr, Jon Frost, and Leonardo Gambacorta. Crypto shocks and retail losses. In *BIS Bulletin*. Bank for International Settlement, February 2023.

[23] Robert Crosnoe and Clea McNeely. Peer relations, adolescent behavior, and public health research and practice. *Family & Community Health*, 31:S71–S80, 2008.

[24] Anthony Cuthbertson. Student accidentally becomes a millionaire after turning selfies into NFT as a joke. *The Independent*, 2022. `https://www.independent.co.uk/life-style/gadgets-and-tech/nft-cryptocurrency-selfie-crypto-b1996276.html`, as of June 11, 2024.

[25] Shayan Eskandari, David Barrera, Elizabeth Stobert, and Jeremy Clark. A First Look at the Usability of Bitcoin Key Management. *NDSS Workshop on Usable Security (USEC)*, 2015.

[26] Julie Ryan Evans. Tales From the Crypto Crypt: 38% of Investors Have Lost More Money Than Made It. Technical report, Lending Tree, January 2023.

[27] Casey Fiesler and Nicholas Proferes. "Participant" Perceptions of Twitter Research Ethics. *Social Media + Society*, 4(1):2056305118763366, 2018.

[28] FINRA. Crypto assets. `https://www.finra.org/rules-guidance/key-topics/crypto-assets`, as of June 11, 2024.

[29] Brock Floyd, Jared Jackson, Emma Probst, Hong Liu, Nischal Mishra, and Chen Zhong. Understanding Learners' Interests in Cybersecurity Competitions on Reddit. In *Proceedings of the 13th International Conference on Education Technology and Computers*, ICETC '21, page 444–449, New York, NY, USA, 2022. Association for Computing Machinery.

[30] Society for Worldwide Interbank Financial Telecommunication. What is KYC?, 2024. `https://www.swift.com/your-needs/financial-crime-cyber-security/know-your-customer-kyc/meaning-kyc`, as of June 11, 2024.

[31] Michael Fröhlich, Felix Gutjahr, and Florian Alt. Don't lose your coin! Investigating Security Practices of Cryptocurrency Users. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, pages 1751–1763, 2020.

[32] Roland Gemayel and Alex Preda. Performance and learning in an ambiguous environment: A study of cryptocurrency traders. *International Review of Financial Analysis*, 77:101847, 2021.

[33] Diptiben Ghelani. What is Non-fungible token (NFT)? A short discussion about NFT Terms used in NFT. *Authorea Preprints*, 2022.

[34] Maria Glenski, Corey Pennycuff, and Tim Weninger. Consumers and Curators: Browsing and Voting Patterns on Reddit. *IEEE Transactions on Computational Social Systems*, 4(4):196–206, 2017.

[35] Andreas Hackethal, Tobin Hanspal, Dominique M Lammer, and Kevin Rink. The Characteristics and Portfolio Behavior of Bitcoin Investors: Evidence from Indirect Cryptocurrency Investments*. *Review of Finance*, 26(4):855–898, December 2021.

[36] M Todd Henderson and Max Raskin. A regulatory classification of digital assets: toward an operational howey test for cryptocurrencies, icos, and other digital assets. *Colum. Bus. L. Rev.*, page 443, 2019.

[37] Joe Hernandez and Jonathan Franklin. The SEC charges Lindsay Lohan, Jake Paul and others with illegally promoting crypto. *NPR*, March 2023. `https://www.npr.org/2023/03/22/1165477713/lindsay-lohan-jake-paul-sec-crypto`, as of June 11, 2024.

[38] Rebecca Hersher. Meet Mafiaboy, The 'Bratty Kid' Who Took Down The Internet. *NPR*, 2015. `https://www.npr.org/sections/alltechconsidered/2015/02/07/384567322/meet-mafiaboy-the-bratty-kid-who-took-down-the-internet`, as of June 11, 2024.

[39] Stan Higgins. Bitcoin Tipping Service ChangeTip to Shut Down. *CoinDesk*, September 2016. `https://www.coindesk.com/markets/2016/11/18/bitcoin-tipping-service-changetip-to-shut-down/`, as of June 11, 2024.

[40] Liselot Hudders, Steffi De Jans, and Marijke De Veirman. The commercialization of social media stars: a literature review and conceptual framework on the strategic use of social media influencers. *International Journal of Advertising*, 40(3):327–375, 2021.

[41] Eaman Jahani, Peter M. Krafft, Yoshihiko Suhara, Esteban Moro, and Alex Sandy Pentland. ScamCoins, S*** Posters, and the Search for the Next BitcoinTM: Collective Sensemaking in Cryptocurrency Discussions. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW), November 2018.

[42] Haiyan Jia, Pamela J. Wisniewski, Heng Xu, Mary Beth Rosson, and John M. Carroll. Risk-taking as a learning process for shaping teen's online information privacy behaviors. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, page 583–599, New York, NY, USA, 2015. Association for Computing Machinery.

[43] Benjamin Johnson, Steven Co, Tianze Sun, Carmen CW Lim, Daniel Stjepanović, Janni Leung, John B Saunders, and Gary CK Chan. Cryptocurrency trading and its associations with gambling and mental health: A scoping review. *Addictive Behaviors*, 136:107504, 2023.

[44] Benjamin Johnson, Daniel Stjepanović, Tianze Sun Janni Leung, and Gary C. K. Chan. Cryptocurrency trading, mental health and addiction: a qualitative analysis of reddit discussions. *Addiction Research & Theory*, 31, 2023.

[45] Daisuke Kawai, Alejandro Cuevas, Bryan Routledge, Kyle Soska, Ariel Zetlin-Jones, and Nicolas Christin. Is your digital neighbor a reliable investment advisor? In *Proceedings of the 32nd Web Conference (WWW'23)*, pages 3581–3591, Austin, TX, May 2023.

[46] Megan Knittel, Shelby Pitts, and Rick Wash. "The Most Trustworthy Coin"': How Ideological Tensions Drive Trust in Bitcoin. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019.

[47] Katharina Krombholz, Aljosha Judmayer, Matthias Gusenbauer, and Edgar Weippl. The other side of the coin: User experiences with bitcoin security and privacy. In *Financial Cryptography and Data Security: 20th International Conference, FC 2016, Christ Church, Barbados, February 22–26, 2016, Revised Selected Papers 20*, pages 555–580. Springer, 2017.

[48] Adena B Meyers Laura E Berk. *Infants, Children, and Adolescents*. Pearson, 8 edition, 2016.

[49] Van Thanh Le. How to Buy Crypto Under 18: Your Ultimate Guide to Financial Freedom. *Coin360*, October 2023. `https://coin360.com/news/how-to-buy-crypto-under-18`, as of June 11, 2024.

[50] Sonia Livingstone, Mariya Stoilova, and Rishita Nandagiri. Children's data and privacy online: growing up in a digital age: an evidence review. *LSE Research Online*, 2019.

[51] Elizabeth Lopatto. Binance really loved telling people to use vpns, allegedly. *The Verge*, March 2023. `https://www.theverge.com/2023/3/27/23659109/binance-cftc-vpn-signal-notes`, as of June 11, 2024.

[52] Alexandra Mai, Katharina Pfeffer, Matthias Gusenbauer, Edgar Weippl, and Katharina Krombholz. User Mental Models of Cryptocurrency Systems - A Grounded Theory Approach. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*, pages 341–358. USENIX Association, August 2020.

[53] Shaurya Malwa. Reddit Starts Airdrop of Polygon-Based 'Collectible Avatars'. *CoinDesk*, August 2022. `https://www.coindesk.com/business/2022/08/25/reddit-starts-airdrop-of-polygon-based-collectible-avatars/`, as of June 11, 2024.

[54] Prensky Marc. Digital natives, digital immigrants. *On the horizon*, 9(5):1–6, 2001.

[55] Alice E Marwick and danah boyd. Networked privacy: How teenagers negotiate context in social media. *New Media & Society*, 16(7):1051–1067, 2014.

[56] Sarah Meiklejohn, Marjori Pomarole, Grant Jordan, Kirill Levchenko, Damon McCoy, Geoffrey M Voelker, and Stefan Savage. A fistful of Bitcoins: characterizing payments among men with no names. In *Proceedings of the ACM/USENIX Internet measurement conference*, pages 127–140, Barcelona, Spain, October 2013.

[57] Julio C. Mendoza-Tello, Higinio Mora, Francisco A. Pujol-López, and Miltiadis D. Lytras. Social Commerce as a Driver to Enhance Trust and Intention to Use Cryptocurrencies for Electronic Payments. *IEEE Access*, 6:50737–50751, 2018.

[58] Jonathan Meng and Feng Fu. Understanding gambling behaviour and risk attitudes using cryptocurrency-based casino blockchain data. *Royal Society open science*, 7(10):201446, 2020.

[59] Mehrnoosh Mirtaheri, Sami Abu-El-Haija, Fred Morstatter, Greg Ver Steeg, and Aram Galstyan. Identifying and analyzing cryptocurrency manipulations in social media. *IEEE Transactions on Computational Social Systems*, 8(3):607–617, 2021.

[60] Ujan Mukhopadhyay, Anthony Skjellum, Oluwakemi Hambolu, Jon Oakley, Lu Yu, and Richard Brooks. A brief survey of Cryptocurrency systems. In *2016 14th Annual Conference on Privacy, Security and Trust (PST)*, pages 745–752, 2016.

[61] Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system, 2008. `https://bitcoin.org/en/bitcoin-paper`, as of June 11, 2024.

[62] Satoshi Nakamoto. Bitcoin open source implementation of P2P currency, February 2009. `https://www.bitcoin.com/satoshi-archive/forum/p2p-foundation/1/#selection-4.0-4.3`, as of June 11, 2024.

[63] Philip WS Newall and Leonardo Weiss-Cohen. The gamblification of investing: How a new generation of investors is being born to lose. *International Journal of Environmental Research and Public Health*, 19(9):5391, 2022.

[64] James Nicholson, Yousra Javed, Matt Dixon, Lynne Coventry, Opeyemi Dele Ajayi, and Philip Anderson. Investigating teenagers' ability to detect phishing messages. In *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 140–149. IEEE, 2020.

[65] Leonardo Nizzoli, Serena Tardelli, Marco Avvenuti, Stefano Cresci, Maurizio Tesconi, and Emilio Ferrara. Charting the Landscape of Online Cryptocurrency Manipulation. *IEEE Access*, 8:113230–113245, 2020.

[66] U.S. Federal Bureau of Investigation. Internet crime report. Technical report, FBI Internet Crime Complaint Center (IC3), March 2022. `https://www.ic3.gov/Media/PDF/AnnualReport/2022_IC3Report.pdf`, as of June 11, 2024.

[67] U.S. Federal Bureau of Investigation. Elder fraud, 2023. `https://www.fbi.gov/how-we-can-help-you/scams-and-safety/common-scams-and-crimes/elder-fraud`, as of June 11, 2024.

[68] U.S. Department of Justice. Bitfinex Hacker and Wife Plead Guilty to Money Laundering Conspiracy Involving Billions in Cryptocurrency, August 2023. `https://www.justice.gov/opa/pr/bitfinex-hacker-and-wife-plead-guilty-money-laundering-conspiracy-involving-billions`, as of June 11, 2024.

[69] Sebeom Oh, Samuel Rosen, and Anthony Lee Zhang. Digital Veblen Goods. *SSRN*, December 2023. `https://ssrn.com/abstract=4042901`.

[70] Kostantinos Papadamou, Jay Patel, Jeremy Blackburn, Philipp Jovanovic, and Emiliano De Cristofaro. From HODL to MOON: Understanding Community Evolution, Emotional Dynamics, and Price Interplay in the Cryptocurrency Ecosystem. *arXiv*, 2023. `https://arxiv.org/abs/2312.08394`.

[71] Ross C. Phillips and Denise Gorse. Mutual-Excitation of Cryptocurrency Market Returns and Social Media Topics. In *Proceedings of the 4th International Conference on Frontiers of Educational Technologies*, ICFET '18, page 80–86, New York, NY, USA, 2018. Association for Computing Machinery.

[72] Nathaniel Popper. Lost passwords lock millionaires out of their Bitcoin fortunes. *The New York Times*, 12, 2021. `https://www.nytimes.com/2021/01/12/technology/bitcoin-passwords-wallets-fortunes.html`, as of June 11, 2024.

[73] Nicholas Proferes, Naiyan Jones, Sarah Gilbert, Casey Fiesler, and Michael Zimmer. Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics. *Social Media + Society*, 7(2):20563051211019004, 2021.

[74] Ivana Rakovic and Yavuz Inal. Dark finance: exploring deceptive design in investment apps. In *IFIP Conference on Human-Computer Interaction*, pages 339–348. Springer, 2023.

[75] Reddit recap 2021, December 2021. `https://www.redditinc.com/blog/reddit-recap-2021`, as of June 11, 2024.

[76] Evan Ryan. Save the kids: The need for regulation of cryptocurrency to protect adolescents from fraud. *Family Court Review*, 2023.

[77] U.S. Securities and Exchange Commision. SEC Charges Coinbase for Operating as an Unregistered Securities Exchange, Broker, and Clearing Agency, June 2023. `https://www.sec.gov/news/press-release/2023-102`, as of June 11, 2024.

[78] U.S. Securities and Exchange Commision. SEC Charges Crypto Company SafeMoon and its Executive Team for Fraud and Unregistered Offering of Crypto Securities, November 2023. `https://www.sec.gov/news/press-release/2023-229`, as of June 11, 2024.

[79] U.S. Securities and Exchange Commision. SEC Charges Kraken for Operating as an Unregistered Securities Exchange, Broker, Dealer, and Clearing Agency, November 2023. `https://www.sec.gov/news/press-release/2023-237`, as of June 11, 2024.

[80] Kyle Soska, Jin-Dong Dong, Alex Khodaverdian, Ariel Zetlin-Jones, Bryan Routledge, and Nicolas Christin. Towards understanding cryptocurrency derivatives: A

case study of BitMEX. In *Proceedings of the 30th Web Conference (WWW'21)*, Ljubljana, Slovenia (online), April 2021.

[81] Fred Steinmetz, Marc von Meduna, Lennart Ante, and Ingo Fiedler. Ownership, uses and perceptions of cryptocurrency: Results from a population survey. *Technological Forecasting and Social Change*, 173:121073, 2021.

[82] Hannah R Stevens, Irena Acic, and Sofia Rhea. Natural language processing insight into LGBTQ+ youth mental health during the COVID-19 pandemic: Longitudinal content analysis of anxiety-provoking topics and trends in emotion in LGBTeens microcommunity subreddit. *JMIR public health and surveillance*, 7(8):e29029, 2021.

[83] Chris Stokel-Walker. How a Squid Game Crypto Scam Got Away With Millions. *WIRED*, 2021. https://www.wired.com/story/squid-game-coin-crypto-scam/, as of June 11, 2024.

[84] Sam Taube. What Is a Custodial Account? UGMAs, UTMAs and More. *NerdWallet*, March 2024. https://www.nerdwallet.com/article/investing/what-is-a-custodial-account , as of June 11, 2024.

[85] Mary Theofanos, Yee-Yin Choong, and Olivia Murphy. Passwords Keep Me Safe' – Understanding What Children Think about Passwords. In *USENIX Security Symposium*, pages 19–35, 2021.

[86] Taro Tsuchiya, Alejandro Cuevas, Thomas Magelinski, and Nicolas Christin. Misbehavior and Account Suspension in an Online Financial Communication Platform. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 2686–2697, New York, NY, USA, 2023. Association for Computing Machinery.

[87] Tejaswi Volety, Shalabh Saini, Thomas McGhin, Charles Zhechao Liu, and Kim-Kwang Raymond Choo. Cracking Bitcoin wallets: I want what you have in the wallets. *Future Generation Computer Systems*, 91:136–143, 2019.

[88] Artemij Voskobojnikov, Oliver Wiese, Masoud Mehrabi Koushki, Volker Roth, and Konstantin (Kosta) Beznosov. The U in Crypto Stands for Usable: An Empirical Study of User Experience with Mobile Cryptocurrency Wallets. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[89] Kristin Weber, Andreas E Schütz, Tobias Fertig, and Nicholas H Müller. Exploiting the human factor: Social engineering attacks on cryptocurrency users. In *Learning and Collaboration Technologies. Human and Technology Ecosystems: 7th International Conference, LCT 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part II 22*, pages 650–668. Springer, 2020.

[90] Helen Whittle, Catherine Hamilton-Giachritsis, Anthony Beech, and Guy Collings. A review of online grooming: Characteristics and concerns. *Aggression and Violent Behavior*, 18(1):62–70, 2013.

[91] Stephen Wooley, Andrew Edmonds, Arunkumar Bagavathi, and Siddharth Krishnan. Extracting Cryptocurrency Price Movements from the Reddit Network Sentiment. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 500–505, 2019.

[92] Qunfang Wu, Louisa Kayah Williams, Ellen Simpson, and Bryan Semaan. Conversations About Crime: Re-Enforcing and Fighting Against Platformed Racism on Reddit. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW1), April 2022.

[93] Saijun Zhang, Meirong Liu, Yeefay Li, and Jae Eun Chung. Teens' social media engagement during the covid-19 pandemic: a time series examination of posting and emotion on reddit. *International Journal of Environmental Research and Public Health*, 18(19):10079, 2021.

[94] Chengyan Zhu, Shiqing Huang, Richard Evans, and Wei Zhang. Cyberbullying among adolescents and children: A comprehensive review of the global situation, risk factors, and preventive measures. *Frontiers in public health*, 9:634909, 2021.

[95] Şahin Telli and Hongzhuan Chen. Multifractal behavior relationship between crypto markets and Wikipedia-Reddit online platforms. *Chaos, Solitons & Fractals*, 152:111331, 2021.

A   Code Book

The following lists the codes derived from the inductive analysis of Reddit posts. Each definition includes an example of one of the posts assigned this code, the number of posts assigned this code, and the number of deleted or removed posts that were assigned this code.

Irrelevant: Not actually discussing cryptocurrency, but containing a keyword. e.g., ***Title:*** *We ZOINKED the kraken! Haha!* ***Body:*** *The kraken is now dead, ZOINKED by our fine crew.But the JINKIES nation lies in the distance. Prepare for D-Day, and prepare to ***ZOINK****

Count: 181 (10.8% of sample), 16 removed/deleted

Other (relevant): Post that mentions cryptocurrency but does not fit into other categories and/or does not really have content of interest. e.g., *Title: Fun fact:* **Body:** *In germany it's legal to scam nfts, because they don't count as reallife possessions, same with cs:go skins etc.*
Count: 19 (1.1% of sample), 3 removed/deleted

Irrelevant, cryptocurrency mentioned in passing: Cryptocurrency is mentioned in passing but is not actually the subject of the post or discussion in the comments. e.g., **Title:** *After you guys complete school what will you do?* **Body:** *everyone: become a gamer and and nft bro [newline] no like fr what will you do?*
Count: 86 (5.1% of sample), 3 removed/deleted

Deleted/Removed (ambiguous content): Post content is removed or deleted. The title mentions cryptocurrency, but it's impossible to determine exactly what the content was about. There are either no comments or the existing comments do not permit us to infer the content. e.g., **Title:** *Cryptocurrency* **Body:** *[removed]*
Count: 67 (4.0% of sample), all removed/deleted

Joke or meme: Post is either explicitly labeled as or is clearly interpreted by users as a joke; it also applies to meme images or videos. e.g., **Title:** *You think its FUNNY to take screenshots of people NFT huh* **Body:** *Property theft is a joke to you*
Count: 254 (15.2% of sample), 29 removed/deleted

Sarcasm: The post is not clearly a joke, but the content is so outrageous that it seems likely that it is sarcastic. e.g., **Title:** *GIVEAWAY ALERT Will send 1 Bitcoin to everyone who sends me 2 bitcoins HURRY!* **Body:** *Edit:this is no scam trust me bro*
Count: 52 (3.1% of sample), 16 removed/deleted

Reddit NFT: Post mentions the Reddit NFTs. e.g., **Title:** *Nft update, accepted it and it looks terrible, thank you Reddit* **Body**: *[Post linked to image of Reddit NFT]*
Count: 156 (9.3%), 28 removed/deleted

Criticism of crypto: Post is either a critique of crypto in general or specific projects. e.g., **Title:** *NFTS are scams and not worth any sort of money* **Body:** *You are buying the rights to some gettyimages monkey for $200k lol*
Count: 110 (6.6% of sample), 24 removed/deleted

Discussion of obtaining/holding crypto: Post includes a discussion of purchasing or otherwise obtaining crypto, including via giveaways; also includes situations where users just describe that they own crypto. e.g., **Title:** *I put my entire allowence into crypto* **Body:** *[deleted]*

Count: 226 (13.5% of sample), 57 removed/deleted

Discussion of mining crypto: Posts discusses mining crypto currency. e.g., **Title:** *[other] Late night Dogecoin miner thread, go!* **Body:** *[Post linked to an image of a terminal running Dogecoin mining software]*
Count: 59 (3.5% of sample), 21 removed/deleted

Discussion of trading crypto: Post discusses trading crypto currency or NFTs. Includes instances where users describe the act of trading without calling it trading. e.g., **Title:** *i'm trading 10$ worth of crypto* **Body:** *so quirky amirite [newline] no*
Count: 25 (1.5% of sample), 6 removed/deleted

Discussion of transferring crypto: Comments or post include discussion of users selling or otherwise transferring crypto. e.g., **Title:** *found some emails from 2017 showing Bitcoin i sent/received. This is now worth $333* **Body:** *[Post linked to an image showing a Coinbase confirmation email for sending 0.008931 Bitcoin]*
Count: 30 (1.8% of sample) , 15 removed/deleted

Cryptocurrency expected to increase in value: Post suggests an expected boost in a cryptocurrency or other digital asset. e.g., **Title:** *DOGECOIN TO THE MOON* **Body:** *[removed]*
Count: 29 (1.7% of sample), 18 removed/deleted

Profit: Post includes discussion of user making money in some way off of crypto assets. e.g., **Title:** *Bought more of a crypto coin then it went up 10% within a few hours lul.* **Body:***[removed]*
Count: 51 (3.0%), 13 removed/deleted

Loss: Post includes discussion of a user losing money in some way off of NFTs or another crypto. e.g., **Title:** *Lost almost 5K in my crypto since Elon Musk was a Douche on twitter and value plummeted by over 30%. I feel so dispondant. I need hugs.* **Body:** *[Post linked to an image of a portfolio of crypto assets worth around $7,700.]*
Count: 17 (1.0% of sample), 6 removed/deleted

Explicit promotion of specific project: Posts encourage users to use a particular project or service. e.g., **Title:** *Pi Network Cryptocurrency the most promising Cryptocurrency.* **Body:** *[removed]*
Count: 305 (18.2% of sample) , 228 removed/deleted

General promotion of crypto: The post generally promotes crpyto assets without specifying a specific project; includes posts that push back against criticism. e.g., **Title:** *Cryptocurrency is the only valuable thing in the world that can be set*

up in a way where nobody can take it from you. **Body:** *[no text]*
Count: 35 (2.1% of sample) , 8 removed/deleted

Informational question about crypto: Post or comments that asks for information about crypto (e.g., what it is, how it works, etc.) e.g., ***Title:*** *Can someone explain to me what is dogecoin is it a new form of crypto currency or sumthin.* ***Body:*** *[deleted]*
Count: 68 (4.1% of sample), 11 removed/deleted

Subjective question about crypto: Post that asks a subjective question about users' perspective on cryptocurrency or personal experience with cryptocurrency. e.g., ***Title:*** *What is your opinion on NFTs?* ***Body:*** *I'm curious to hear what everyone thinks about these things. edit: Interesting. It sounds like most people don't like them, but also dont know very much about them.*
Count: 105 (6.3% of sample), 34 removed/deleted

Using crypto for payment: Post mentions or is about using crypto for payment. e.g., ***Title:*** *SKIDS ACCOUNTS SHOP QUALITY SUPPORT + LIFETIME Warranties - ONLY $4+ Netflix, Hulu, Spotify, etc. [W] PayPal, BTC, Cash App.* ***Body:*** *[removed]*
Count: 25 (1.5% of sample), 10 removed/deleted

Crime: Post or comments suggest that the user has fallen victim to a scam, virus, or other crime; also applies to posts that describe a crime, even if the user is not the victim. e.g., ***Title:*** *Crypto scammers are the scum of the earth.* ***Body:*** *So, i went onto my YouTube subscriptions inbox today, only to notice a Livestream promoting Ethereum. Turns out one of the YT channels I'm subbed to was hacked by crypto scammers. Seriously, fuck these people. I know it's fun to laugh at NFT collectors, spending millions on Bored Apes, but screw these people.*
Count: 27 (1.6% of sample) , 4 removed/deleted

Working on a crypto project: Post or comments discuss a teenager creating or working for crypto asset project. e.g., ***Title:*** *I made some new nfts* ***Body:*** *So far I have giga chad,*

peter Griffin, moist critical, and American syco. Who else should I make?
Count: 36 (2.1% of sample), 3 removed/deleted

Discussion of news in crypto: Post or comments about news related to a crypto project. e.g., ***Title:*** *Binance Acquired regulated Crypto exchange.* ***Body:*** *[Post linked to article about acquisition]*
Count: 69 (4.1% of sample), 13 removed/deleted

Giveaway: Post or comment purports to be giving away/airdropping some sort of cryptocurrency; includes "faucets." e.g., ***Title:*** *Free Bitcoin for Teens Only.* ***Body:*** *Make any comment below for your complimentary 100 bits of Bitcoin and it will be delivered to you via the ChangeTip bot! What is Bitcoin? What is ChangeTip? Edit: I am going to bed, but leave your comments below if you haven't been tipped and I will get to you in the morning. Night.*
Count: 145 (8.7% of sample) , 118 removed/deleted

Gambling: Post discusses gambling with crypto. e.g., ***Title:*** *I just made $250 gambling bitcoin.* ***Body:*** *[Post links to a referral code for Bitsler.com]*
Count: 13 (0.8% of sample), 3 removed/deleted

Begging: Post requests that users send cryptocurrency to the poster in exchange for no consideration. e.g., ***Title:*** *Anyone have some bitcoin to spare.* ***Body:*** *[removed]*
Count: 8 (0.5% of sample), 5 removed/deleted

B Filtering Keywords

The following tables show the regular expression strings used to identify posts that were potentially about crypto in the teenage subreddits. Matches were identified using Python 3.9.7 re package. The \b at the beginning and end of each string ensures that matches are only returned if the match is a free-standing word. For example, \bterra\b matches "terra" or "(terra)" but not "terrarium." Complete documentation can be found at https://docs.python.org/3.9/library/re.html.

Table 3: Keywords derived from the top ten coins by market cap, according to `coinmarketcap.com` as of June 1st, 2023. For most currencies, regex strings are included for both the currency's name and abbreviation(s). Some names and abbreviations are excluded, as we anticipated that their inclusion cause too many false positives (i.e., polygon, ripple, and doge). A regex string for "binance coin" was also excluded, since there is a string for "binance" in another category of keywords

Keyword	# in Sample	% Irrelevant (#)
\bbit(|-)?coin(s)?\b	280	7.5% (21)
\bbtc\b	33	3.0% (1)
\bdoge(|-)?coin(s)?\b	83	6.0% (5)
\bethereum(s)?\b	56	3.6% (2)
\beth\b	35	60.0% (21)
\bether(s)?\b	removed	removed
\btether(s)?\b	14	85.7% (12)
\busdt\b	3	33.3% (1)
\bmatic\b	5	100.0% (5)
\bsolana\b	3	0.0% (0)
\bsol\b	removed	removed
\bcardano\b	1	0.0% (0)
\bada\b	removed	removed
\bxrp\b	1	0.0% (0)
\busd(|-)?coin(s)?\b	2	50.0% (1)
\busdc\b	1	100.0% (1)
\bbnb\b	6	83.3% (5)

Table 4: Keywords derived from three high-profile failed crypto projects: the Terra USD stable coin [13], SafeMoon [78], and Squid Game Token [83]. For TerraUSD, we include both the name of the project (Terra) and the abbreviations for the pair of coins that were core to the algorithmic stablecoin system (TerraUSD and Luna). For Squid Game Token, we include two possible variations (Coin or Token)

Keyword	# in Sample	% Irrelevant (#)
\blun(c)?\b	4	100.0% (4)
\bust(c)?\b	33	93.9% (31)
\bterra\b	34	97.1% (33)
\bsafe(|-)?moon\b	8	0.0% (0)
\bsquid(|-)?game(|-)?coin\b	0	—
\bsquid(|-)?game(|-)?token\b	0	—

Table 5: Keywords derived from the names of centralized crypto exchanges, decentralized crypto exchanges, and interest bearing services.

Keyword	# in Sample	% Irrelevant (#)
\bbinance\b	7	0.0% (0)
\bftx\b	3	0.0% (0)
\bcoin(|-)?base\b	10	10.0% (1)
\bkraken\b	35	100.0% (35)
\bsushi(|-)?swap\b	0	—
\buni(|-)?swap\b	2	0.0% (0)
\bpancake(|-)?swap\b	2	0.0% (0)
\bcelsius\b	removed	removed
\bblock(|-)?fi\b	0	—
\bnexo\b	3	100.0% (3)

Table 6: Keywords derived from the names of cryptocurrency gambling services

Keyword	# in Sample	% Irrelevant (#)
\bstake(\.com|\.us)\b	1	0.0% (0)
\bcloud(|-)?bet\b	0	—
\bmeta(|-)?spin(s)?(|-)?casino\b	0	—
\b7(|-)?bit(|-)?casino\b	0	—
\bbets\.io\b	0	—
\bbit(-|)?starz\b	0	—
\bm(-|)?bit(|-)?casino\b	0	—
\broll(|-)?bit\b	0	—

Table 7: Keywords derived from words and concepts related to crypto assets.

Keyword	# in Sample	% Irrelevant (#)
\bcrypto(s)?\b	281	11.0% (31)
\bblock(|-)?chain(s)?\b	35	25.7% (9)
\bnft(s)?\b	680	3.5% (24)
\bdefi\b	5	20.0% (1)
\bcrypto(|-)?currenc(y|ies)\b	173	12.1% (21)
\bdao(s)?\b	1	100.0% (1)
\bdecentralize(d|s)?\b	12	66.7% (8)
\bstable(|-)?coin(s)?\b	1	0.0% (0)
\bmoon(|-)?shot(s)?\b	8	100.0% (8)
\bdegen(s)?\b	removed	removed
\baping\b	removed	removed
\bde(|-)?peg(s|ging)?\b	0	—
\bshill(s|ing|ed)?\b	removed	removed
\bmarket(|-)?cap(s)?\b	8	12.5% (1)

"I can say I'm John Travolta … but I'm not John Travolta." *Investigating the Impact of Changes to Social Media Verification Policies on User Perceptions of Verified Accounts

Carson Powers*, Nickolas Gravel*, Christopher Pellegrini*, Micah Sherr**
Michelle L. Mazurek[†], and Daniel Votipka*
*Tufts University; [†]University of Maryland; **Georgetown University
{carson.powers,nickolas.gravel,christopher.pellegrini,daniel.votipka}@tufts.edu
mmazurek@cs.umd.edu; msherr@cs.georgetown.edu

Abstract

Until recently, almost all social media platforms verified the identities behind notable accounts. Prior work showed users understood this process. However, Twitter/𝕏's switch to an open, less rigorous verification process represented a significant policy shift. We conduct a U.S. Census-representative survey to investigate how this and subsequent verification changes across social media impact users' verification perceptions. We find most users generally recognize the changes to Twitter/𝕏's policy, though many still believe Twitter/𝕏 verifies account holders' true identities. However, users are less aware of subsequent Facebook verification changes. We also find platforms' verification differences do not impact user perceptions of posted content credibility.

Finally, we investigate hypothetical verification policies. We find participants are more likely to perceive posts from verified accounts as credible when only notable accounts are eligible and government document review is required. Payment did not affect credibility decisions, but participants felt strongly that payment for verification was unacceptable.

1 Introduction

Most social media sites, such as Twitter/𝕏[1], Facebook, Tik-Tok, and LinkedIn, support some form of account verification. Each platform reviews accounts [42], then adds a badge (e.g.,

---

*The full quote by a participant asked what verification policy changes they would suggest was, "I would require a photo ID. I can say I'm John Travolta and I can give you my email address (which can be almost anything) to confirm me, but I'm not John Travolta."

[1]Since Twitter's rebranding to 𝕏 occurred after our survey, we will use "Twitter" in the remainder of the paper.

✔) next to the *verified account's* (VA)[2] username to signal the verification process has been completed. VAs were introduced to help users differentiate between accounts belonging to the entity named (often a celebrity or account of public interest) and parodies or impostors [74]. Twitter introduced VAs in 2009 following a rise in impostor accounts [67], and other platforms followed suit [23, 34, 42, 62, 64, 72]. With the rise of disinformation on social media, the value of determining a post's true source is growing [29, 36–38, 49, 59]. This challenge is exacerbated in emergencies, when users look to social media for real-time information [7, 32, 41, 47, 76]. During terrorist and active-shooter events [4, 7] and natural disasters [41, 46, 54], users look to local authorities, such as police and fire departments, for safety information. Without rigorous account verification, users may trust false information with life threatening consequences [28].

While there is some evidence suggesting users equate account verification with credibility [44], other work has shown, in isolation, users correctly understood the verification badge only indicated authenticity [75]. However, recent changes to verification policies may muddle verification's purpose. First, the social media ecosystem has splintered, with new and niche platforms growing (e.g., TikTok, Truth Social, etc.). While verification is similar across platforms, some subtle differences should impact the correct interpretation of VAs.

Additionally, some of the largest existing platforms have made significant policy changes. Most notably, Twitter dramatically changed its verification policy after being acquired by Elon Musk in October 2022. Prior to the purchase, Twitter verified notable users' accounts (e.g., celebrities and public figures or organizations) by requiring proof of identity via a government-issued ID [74]. Twitter then made verification available to any user for a monthly $8[3] subscription fee, and swapped government ID for a verified phone number [73]. This transition was tumultuous, with abrupt changes regularly covered in the media [10, 21, 30, 33, 45, 48]. Some users

---

[2]Terms differ by platform. For consistency, we refer to accounts that have undergone some form of authentication as verified accounts (VAs).

[3]$12 if signing up in-app to account for Apple's/Google's service charges.

took advantage of the new policy to establish impostor accounts [50, 66]. To less fanfare, Facebook also adjusted its verification policy, allowing anyone to obtain a VA for a fee, but maintaining the requirement for ID verification, and LinkedIn made verification slightly more open without adding a fee.

We seek to assess the impact of these policy changes on user perceptions of VAs on Facebook and Twitter, as well as how users think verification policies *should* work. Towards that goal, this paper considers three research questions:

RQ1: What are the verification policies used by popular social media platforms and how have they changed over time?

RQ2: What do users think account verification entails? How does it impact perceptions of posted content credibility?

RQ3: How would potential changes to verification policies impact user perceptions of posts from verified accounts and user perceptions of the policies?

RQ1 seeks to understand the VA ecosystem. Due to the fractured landscape, perceptions may vary depending on the platforms used. With the volume of media coverage and rapid policy-making during the Twitter transition, user perceptions may represent a snapshot in time, rather than an accurate depiction of current policy. To understand the impact of these changes, we must first enumerate verification policies.

To address RQ1, we captured the verification policies of eight popular social media sites from April 2022 to August 2023, noting any changes. After enumerating verification policies, we conducted a controlled experiment—using a vignette-based survey of 1600 U.S. Prolific users—to address RQ2 and RQ3 for a U.S. population using text-based social media. Participants were first shown two mock posts containing contradictory information and asked to indicate which they perceived as more credible, to test the VA's impact on their assessment of relative credibility when presented with information from similar accounts—a common challenge when assessing information during emergency events. We varied the platform (Twitter vs. Facebook) and asked participants to indicate how they believed their assigned platform defined verification. Then, we presented participants with a new verification policy and asked them to reevaluate the previously shown mock posts with this new policy in mind. We also asked participants their perceptions of the new policy.

Participants' understanding of Facebook's and Twitter's verification policies was mixed, and they were more likely to correctly perceive Facebook's policy as requiring identity verification. Participants correctly indicated Twitter's policy was open to anyone for a fee. This seems to indicate users have better understood the Twitter policy over time, compared to a similar survey conducted earlier by Xiao et al., which asked participants to identify features of verification [81]. However, participants seemed unaware of Facebook's policy, with many

still believing verification was free and only for notable accounts. This is likely due to the newness of Facebook's policy change and lack of broad media coverage.

We did not observe differences in participants' assessments of posted content credibility between assigned platforms. However, after providing participants with a verification policy, they were more likely to find posts from the VA credible when government ID was required and only notable accounts were verified. Participants also perceived these policies as more acceptable (matching Xiao et al. [81]). This difference between initial assessment and re-assessment after reviewing a verification policy suggests participants do not consider the details of the policy fully when assessing posts from VAs.

Finally, while participants strongly disliked paying for verification—corroborating Xiao et al. [81]—payment did not impact participants' credibility decisions before or after reviewing the verification policy. While this indicates verification payment has no direct impact on user assessments of credibility of VAs' posts, the strong dislike of the policy may have downstream impacts that should be considered in future work, especially as several participants reported no longer trusting any verification provided by Twitter.

2 Related Work

Credibility of Online Content. Much work has investigated factors affecting user perception of online content credibility. Wineburg et al. assessed students' ability to judge online source credibility [79]. Fogg et al. found the "design look" of a website impacts perceived credibility [20]. Hilligoss and Rieh found users are more likely to find information legitimate when the source appears "official" [27]. Hassoun et al. performed a qualitative analysis of Gen Z's evaluation of online information, finding three "trust heuristics": credible information was easily accessible, neutral in tone, and "felt right." Their participants reported using number of likes and comments as a form of "crowdsourcing credibility" [25]. This mirrors previous findings that users are more likely to perceive information as credible when they believe others perceive it as credible [9, 19, 22, 27, 69], an effect called the *endorsement heuristic*. Familiarity with a source also increases perceived credibility, known as the *reputation heuristic* [43]. We build on prior work, focusing specifically on social media platforms and the effect of verified indicators.

Verification's Impact on Social Media Post Credibility. The verified indicator's purpose is to affirm an account holder's identity, not signal posted content credibility. However, humans' reliance on trust heuristics may lead to an indirect effect on perceived credibility, which may explain conflicting evidence whether users separate *authenticity* and *credibility*.

Early work by Morris et al. suggested the verified indicator highly impacts users' evaluation of credibility [44]. However, their work asked participants to list features they consider

when deciding if a tweet is credible, which measures the *conscious* impact of verification badges, not the *behavioral* impact. Conversely, Vaidya et al. conducted a large-scale controlled experiment, measuring the verified indicator's effect on participants' perceptions of post credibility. They found users understood verification indicated the account holder was who they said they were, but does not add credibility to the post [75]. Dumas and Stough conducted a consumer-behavior study where participants were shown influencer-posted content. They found consumers associate VAs with celebrity more than credibility [14]. In this paper, we seek to assess whether user perceptions have changed due to changes to social media verification policies and expand beyond Twitter to consider other platforms.

Most similar to our work, Xiao et al. investigated user understanding of verified indicators on Twitter, Facebook, and TikTok in the wake of Twitter Blue [81]. They surveyed social media platforms and identified dimensions of each verification policy. Using these, they surveyed 299 U.S. adults asking their definitions of verification and whether they found Twitter's policy acceptable. They found participants were more likely to indicate payment was required for Twitter as opposed to other platforms, but most continued to *incorrectly* assume Twitter verified identities of users with verified indicators. They observed users disliked Twitter's policy because it does not verify identity and requires payment. We build on this study in several ways. First, we conducted a more in-depth review of social media platforms by investigating Musk's Twitter posts, which provide valuable context, and monitoring policies over a longer period, which captured policy changes by Meta and LinkedIn. Next, capturing a snapshot after Meta's policy changes allows a useful comparison over time between the works. We also measured how verified indicators impact perceptions of post credibility. Finally, we conduct between-subjects comparisons, randomly assigning participants to define verification for specific platforms instead of asking for general definitions, and test several possible policy designs for their impact on post credibility decisions and policy acceptability. This gives us a more nuanced view of the changing landscape of VAs and its impact on user behaviors.

3 Verification Policy Review

To address RQ1, we reviewed verification policy changes across eight popular social media platforms from April 2022 to August 2023. We outline our collection and review process and describe changing landscape of social media verification.

3.1 Data Collection and Analysis

We collected verification policies from seven of the top eight social media platforms Americans reported getting their news from in 2022 [8], i.e., Twitter, Facebook, TikTok, Snapchat, LinkedIn, Instagram, and YouTube. We excluded Reddit, which does not support account verification, but included Truth Social to represent small, niche platforms.

For each platform, we captured the verification policy on April 14, 2022 and all subsequent policy changes until August 25, 2023. April 14 marked Musk's expression of interest in acquiring Twitter. This date is a significant marker for our analysis, as it potentially influenced changes in the verification policy landscape. We monitored platform policies until our final participant completed our survey (see Section 4) to ensure we captured changes that could affect user perceptions. Details about our web scraping process are in Appendix B.

We also manually reviewed all of Musk's personal tweets about Twitter's verification policy during this period. Musk regularly made policy pronouncements publicly, which drove news coverage [33, 65] and may have influenced perceptions.

To identify common themes across verification policies, we performed an inductive thematic analysis, allowing policy dimensions to arise from the data [68]. Two researchers collaboratively reviewed the initial policies for each platform and subsequent changes as they were collected. Codes were then discussed with the full research team until full agreement was reached. Because we only sought to identify themes and do not attempt to use results for quantitative comparison, we did not assess inter-rater reliability [39].

3.2 Results

We observe several independent *dimensions* of social media verification policy: who can be verified (Eligibility), how accounts are verified (Verification Method), whether users pay a fee, requirements to prevent "deception," and required activity history. Table 1 summarizes the reviewed policies, including any changes occurring during our review.

Further, we observe three distinct *time periods* of social media verification policy:

Before Musk's Twitter takeover (Period 1). From the start of our review (April 14, 2022) until Musk's takeover of Twitter (October 27, 2022), the policies of all eight social media platforms were similar. All allowed verification only for "Notable" users (e.g., celebrities, journalists, public figures). They required users provide government documents to prove identity and did not charge for verification. There was some variation in what platforms considered "deceptive." These policies prevent accounts from changing their account information (e.g., username), having usernames similar to other accounts, posting spam, or attempting to manipulate the platform.

Musk acquired Twitter (October 27, 2022; Period 2). Musk made sweeping verification policy changes by introducing Twitter Blue on November 9, 2022. This program opened verification to any user, removed user identity checks, and required payment [73]. Musk argued open verification would improve conversation quality [15] and reduce bots by creating a barrier to entry [16, 17]. These changes faced broad criti-

Platform	Icon	Eligibility	Ver. Meth.	Payment	Non-Deceptive	Active
Twitter [73, 74]	✔	Notable → Open	Gov ID → Phone	Free → Paid	No profile changes,[1] spam, misleading behaviors, or platform manipulation	Active past 30 days
Facebook [42]	✔	Notable → Open	Gov ID	Free → Paid	No profile changes,[1] unique	Prior posting history
Instagram [42]	✔	Notable → Open	Gov ID	Free → Paid	No profile changes,[1] unique	Prior posting history
TikTok [72]	✔	Notable	Gov ID	Free	No profile changes[1]	Logged in past 6 months
Snapchat [62]	★	Notable	Gov ID	Free	No misleading behaviors	Regularly post content
LinkedIn [34]	✔	Notable → Open	Gov ID[2]	Free	No profile changes	-
YouTube [23]	✔	Notable	Gov ID[3]	Free	No profile changes[1]	Regularly post content
Truth Social [64]	✖	Notable	Gov ID	Free	No misleading behaviors	Regularly post content

[1] All platforms restricted VAs from changing their username. Some also prevented changes to other profile data, such as profile photos and bios.
[2] LinkedIn's verification is only available to US users (through the CLEAR ID program) or employees of companies participating in LinkedIn's company email verification or Microsoft's Entra Verified ID programs.
[3] YouTube does not verify documentation by default, but reserves the right to request additional documentation if necessary.

Table 1: Summary of verification policy dimensions and verified indicators per platform. → indicates a change in the policy during our review with the left hand side indicating the policy at the start of our review and the right hand side showing the final policy.

cism [10, 21, 30, 33, 45, 48], and verified impostor accounts quickly appeared [50, 66], indicating the changes did not produce Musk's desired effect [70].

Twitter paused Twitter Blue on November 11, 2022 and reintroduced it on December 12, 2022 with modified eligibility requirements to limit impostors. Specifically, users were required to verify a working phone number and must have been active 30 days before verification.[4] Twitter also introduced government (✔) and company (✔) badges which were only available to organizations fitting these descriptions.

Potentially adding to user confusion, users verified under Twitter's original verification policy (Twitter Legacy) maintained their verified indicator. Verification of Twitter Legacy and Blue accounts was indistinguishable when looking at individual posts. The only distinction was an indicator on the Twitter Legacy accounts' profile pages. The Twitter Legacy policy remained in effect until April 1, 2023 [48].

While not directly related to the verification policy, Twitter also began prioritizing posts by VAs (January 5, 2023) [77]. Twitter argued this was to ensure users are most likely to see "content that is relevant, credible, and safe," implying a link between verification and credibility.

During this period, all other platform policies were stable.

Meta and LinkedIn alter policies (February 20, 2023; Period 3). Meta, the parent company of Facebook and Instagram, announced Meta Verified [42]. Like Twitter Blue, this subscription-based verification program was open to all users and required payment. However, Meta continued to require government ID for verification—the most significant difference between Twitter's and Facebook's final policies.

On April 12, 2023, LinkedIn also opened verification eligibility beyond notable users [34]. LinkedIn began allowing U.S. users to verify their identities through the CLEAR ID

---

[4]The policy initially added a 90-day activity period on November 24, 2022, but this was relaxed to 30 days prior to Twitter Blue's restart.
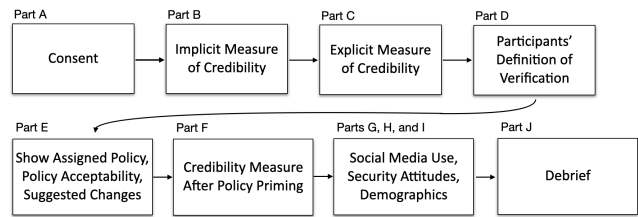
Figure 1: Sections and flow of the user study.

program and verified users with certain corporate email addresses or through the Microsoft Entra Verified company ID program. While not available to all users, it is more open than previously, and follows Meta's example of maintaining identity verification while increasing eligibility.

Our identified dimensions of verification align with those outlined by Xiao et al.'s prior review [81], though our results capture changes to Meta's and LinkedIn's policies that occurred after their review. Our full dataset of policy changes is in supplementary materials [2].

4 Survey Methods

Using the policy dimensions identified in Section 3, we developed an online survey to test participants' understanding of platform policies (RQ2) and their preferences for each policy dimension (RQ3).

4.1 Survey Design

Figure 1 shows the stages of our online survey, which we describe below in turn.

Consent (Part A). We began with a consent form describing the study, potential risks, and data protection procedures. To avoid priming for the verified indicator, which users might
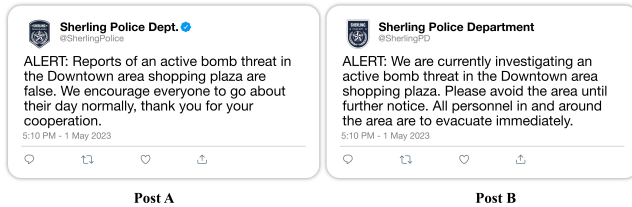
Figure 2: Example Police/Declarative/Twitter condition posts.

otherwise ignore in practice, we used deception when describing the study's purpose, indicating it was to understand how users assess social media posted content credibility.

Implicit effect of verified indicator (Part B, RQ2). Next, participants were shown a pair of posts reporting contradictory information, both from accounts presenting as authorities on the subject. Figure 2 shows an example pair of posts. Posted content details, such as whether they included a verified indicator and the platform for which they were formatted (Twitter or Facebook) varied per condition (see Section 4.2). Participants were asked to indicate which posted content was more likely correct, on a five-point Likert scale. Because the contradictory posts cannot both be true, participants must make some assessment (potentially based on the verified indicator) about account identity to determine which is more credible.

Explicit effect of verified indicator (Part C, RQ2). Next, we asked participants whether the verified indicator affected their posted content credibility choice, on a four-point Likert-type scale from "No effect" to "Major effect." To compare the verified indicator's effect to other account features, participants were asked the same question about the account's picture, name, and handle.[5] The order of account feature questions was randomized to avoid ordering effects [58].

Participants' verification definitions (Part D, RQ2). We then asked participants to define verification to investigate how they understand verification and if this varies by platform.

Assigned verification policy perceptions (Part E, RQ3). We gave a mock verification policy and asked participants to assume their condition-assigned platform adopted this policy. We asked whether they believed it was "acceptable for verifying account owner identity" on a 5-point Likert-type scale from "Unacceptable" to "Acceptable." We also asked them to provide one modification (i.e., addition, deletion, change) to improve the policy. This open-ended question was intended to capture the policy elements participants prefer and prioritize, including those not used on social media platforms.

Credibility perceptions after policy priming (Part F, RQ3). In Part F, we showed participants the original contradictory posts together with their assigned mock verification policy. Then, we repeated Part B's question, asking participants to

---

[5]The account handle question was only included for participants in the Twitter condition because Facebook accounts do not have this feature.

choose which posted content was more credible, this time assuming verification via the given mock policy. Next, we asked participants to assume a friend was unsure which posted content was more credible, and tell us what advice they would give to help the friend decide. This open-ended question captured an additional perspective into participants' credibility assessment. This section included an attention check to identify and remove inattentive respondents [40].

Social media use (Part G), Security attitudes (Part H), and Demographics (Part I). We concluded with questions about our participants' background and demographics. We asked about their social media use for the two platforms tested, as well as more generally. Participants completed Faklaris et al.'s SA-6 scale [18] to assess their computer security practices.

Debrief (Part J). Because we used deception, we debriefed participants about the study's true nature, providing Twitter's and Facebook's verification policies and links to best-practice guidelines for assessing posted content credibility [35,60,61].

4.2 Conditions

Each participant saw two contradictory posts (Parts B and F) and a mock verification policy (Parts E and F). We describe the possible posts and policies defining each *condition*.

Posted content variables. To test the verified indicator's effect, we created four posted content pairs. First, we varied the platform. One of our research questions (**RQ2**) is whether users perceive differences in verification policy between platforms and how this impacts VA credibility perceptions. For this dimension, participants were shown posts using Twitter or Facebook visual cues. This included the posted content design, verified indicator shown (i.e., ✔ vs. ✔), and terminology in survey questions (e.g., "Please answer the following questions considering the two *Twitter* posts above[6]"). We chose these platforms because Twitter changed its verification policy most significantly (see Section 3) and Facebook was the most popular platform with a comparable modality (i.e., YouTube, Instagram, TikTok are mostly image and video-based).

Second, we varied the posted content. Prior work showed content affects users' credibility perceptions [75], so we test multiple content types to avoid bias from a single type.

One pair of posts describes an alleged bomb threat (*Police*), as posted by different accounts (Sherling Police Dept. @*SherlingPolice* or Sherling Police Department @*SherlingPD*) claiming to be the same entity. One post claims the threat is false; the other asserts it is true. The second pair (*Coffee*) appear to be posted by medical doctors (Dr. Samuel Smith, M.D. @*DrSmithMD* or Dr. Alexander Kim, M.D. @*DrKimMD*). The posts contradict about a link between coffee consumption and risk of a disease. Table 2 details the posts. Combining

---

[6]Emphasis not included in survey.

platform and content options produced four posted content conditions. Participants were randomly assigned to one.

The police departments, doctors, and diseases were fictional to eliminate prior knowledge bias. We avoided political topics to prevent polarization effects [31], as prior work showed people distrust evidence contradictory to their beliefs on controversial topics [55]. We chose topics of general importance, where people must rely on expert insights. We chose to use authoritative accounts, as accounts like these could be verified or unverified under all policies reviewed in Section 3, creating a range of reasonable justifications participants could come to in their decision-making. Prior work showed users are more likely to find authoritative accounts credible [75], so we only used authoritative accounts to control for this effect.

To control for potential bias toward declarative or contradictory statements, we randomized which account was verified. We randomized the order the declarative and contradictory posts were shown, to control for ordering effects [58]. To control for other possible credibility indicators, other posted content elements (author profile image, retweets and likes counts, and time since publication) were held constant. Previous research showed these elements significantly affect user perception of posted content credibility [44].

Policy variables. After asking about the pair of posts, we presented participants with mock verification policies to observe how varying policy definitions affect their perception of the verified indicator (**RQ3**). The policies had three variables, representing the three dimensions we observed multiple platforms change in our policy review (Section 3). Table 2b gives the policy text shown for each condition. First, we varied who can be verified (Eligibility). The policy was either *Open*, meaning anyone can apply, or *Notable*, meaning only well-known individuals and organizations are eligible. Next, policies varied in how accounts are verified (Verification Method). That is, accounts must either confirm an email or phone number (*Phone*) or provide government-issued ID (*Gov ID*). Finally, the policy specified whether verification required *Payment*. We used a full-factorial variable combination to create eight policies. Participants were randomly assigned a policy independent of their post and platform condition.

4.3 Recruitment

We conducted our survey on Prolific, a research recruitment service providing high-quality samples [52, 71]. We limited participation to Prolific users at least 18 years old and located in the United States. We used Prolific's census-representative sample feature [56] to ensure a U.S. population-representative distribution by age, gender, and ethnicity. Survey completion time averaged 8.2 minutes, and we paid participants $2.

Content	Position	Posted content Text Summary
Police	Decl.	ALERT: We are currently investigating an active bomb threat in the Downtown area shopping plaza. Please avoid the area...
	Cont.	ALERT: Reports of an active bomb threat in the Downtown area shopping plaza are false...
Coffee	Decl.	Individuals who consume more than three cups of coffee per day may have a higher risk of developing endothrombocytisis.
	Cont.	There have been no research studies that have established a link between coffee consumption and endothrombocytisis.

(a) Posted Content Variables

Dimension	Option	Policy Text
Eligibility	Open	**Any user** can apply for verification
	Notable	Only **well known, high-profile individuals and organizations** can apply for verification
Verification Method	Phone	Accounts are required to **confirm a phone number or email** with the platform
	GovID	Accounts must **submit government-issued identification** matching the name of the account
Payment	Paid	Accounts **pay a monthly subscription fee** to maintain their verification
	Free	Accounts **do not pay any fee** for verification

(b) Policy Variables

Table 2: Summary of (a) posted content and (b) policy conditions. There were four posted content and eight policy conditions, resulting in 32 total conditions after a full-factorial combination.

4.4 Pilot

We piloted the survey with nine participants—drawn from a convenience sample, selected for varying social media familiarity. Pilot participants were asked to "think aloud" while answering questions. We iteratively updated the survey for clarity after each pilot until further changes were unnecessary.

We also tested a third content type about an E.Coli outbreak in lettuce. We recruited 50 participants on Prolific and assigned them randomly to one of the three content types to test whether any content type behaved unexpectedly (e.g., prior experience bias or unexpected relationship with current events). We did not observe unexpected responses, but saw similar results between the E.Coli and Coffee conditions. Therefore, we dropped the E.Coli condition to increase our analysis power by recruiting more participants per condition.

4.5 Data Analysis

Quantitative analysis. To test verification's effect on participants' posted content credibility perceptions before and after stating a policy, and to assess participants' perception of policy acceptability, we used ordinal logistical regressions.

For the two posted content credibility perceptions questions, the outcome variable is a 5-point Likert-scale response regarding which post was correct (Part B and Part F, respectively). Each response was modified to indicate whether the participant perceived the account with or without the verified indicator as correct, to allow for comparisons; e.g., if a participant shown the posts in Figure 2 selected "Definitely A" from the possible options, because A was the VA, their response was modified to "Definitely the VA." For the policy acceptability regression, the outcome variable was the participant's response to the policy acceptability question in Part E.

In each regression, we include the assigned condition's three elements (platform, content, and position) as explanatory variables. For the policy-related regressions (Part E and Part F), we added the policy variables (Eligibility, Verification Method, and Payment). In all regressions, we include demographic explanatory variables (age, gender, education), amount of time spent using Twitter and Facebook, number of social media platforms used, and SA-6 scores. Table 6 in Appendix D summarizes the variables included per regression.

To select a parsimonious model without overfitting, we constructed initial regression models using all possible explanatory variable combinations. We selected models with the minimum Bayesian Information Criterion, appropriate for testing goodness-of-fit [57, 63].

We also examined the explicit impact of verified indicator on credibility perceptions. We compared responses regarding the verified indicator's impact between Twitter- and Facebook-assigned participants using a Pearson's χ^2 test, appropriate for categorical data [51]. Next, we compared responses across the four[7] account features (verified indicator, account username, photo, and handle) using non-parametric, repeated measures tests, appropriate for multiple Likert-scale responses per participant. We began with an omnibus Friedman test across features to control for Type I error; if the result was significant, we applied the Wilcoxon signed-rank test to planned pairwise comparisons of the verified indicator with every other feature [78]. Comparisons were across content conditions.

Qualitative analysis. We used iterative open coding to analyze free-response questions [68]. As our questions were all related to VAs and verification policies, similar to the free-response questions in Vaidya et al. [75], we began with their codebook. However, as verification policies have changed, we allowed additional codes to arise inductively. Three researchers extended the initial codebook collaboratively by reviewing 10 responses. Two researchers independently coded additional responses in rounds of 100, updating the codebook incrementally. After rounds, the coders met, assessed inter-rater reliability using Krippendorff's alpha [26], and resolved coding differences. After two rounds (200 responses), the coders achieved $\alpha = 0.80$, which represents acceptable agree-

ment. The remaining 1386 responses were divided evenly and coded separately by the two coders [26]. Finally, the two researchers performed an axial coding to identify relationships between codes and produce higher-level groups [12, pg. 123-142]. The final codebook is in supplementary materials [2].

To compare initial verification definitions between participants shown Twitter and Facebook posts, we perform Pearson's χ^2 tests, appropriate for categorical data [51]. For each higher-level code group, we compare a code's presence from this group between Twitter- and Facebook-assigned participants. Because this requires multiple testing, we apply a Benjamini-Hochberg correction to adjust p-values [6].

4.6 Ethical Considerations

Tufts University's IRB approved this study. We obtained informed consent prior to the survey. Because we used deception in our study description, we concluded with a debrief and asked participants to re-consent. To avoid response coercion, participants were told they would be paid for completing the survey even if they refused consent, but their response would be deleted. Three participants withdrew after the debriefing.

Responses through Prolific are provided pseudonymously, with only the participant's Prolific ID identifying their response. We did not request additional identifying information.

4.7 Limitations

We presented mock posts, as this provides the control needed to reason about specific variables' effects on credibility perceptions. However, we are unable to capture other credibility perception influences, such as the author's reputation, the viewer's relationship with the author, or viewer's relationships with others who interact with the posted content (e.g., liked or shared). The types of content and other metadata we test are also limited, meaning we are unable to comprehensively test these factors' influence on posted content credibility. We only test textual content, and so our results may not generalize to verified indicators on video-based platforms (e.g., TikTok). We do not test controversial content, as we expect the introduced bias to overwhelm any effect from the verified indicator. This is an inherent tradeoff to limit the study's scope to a reasonable condition set. Our results establish a baseline of verified indicator effect on perceived credibility, and future work should study how the effect changes in the presence of video and controversial topics. We believe our conditions are sufficient to target our study's research questions.

The study's setting also differs from the real world. Participants may have spent more time reviewing our contradictory posts than when casually browsing social media feeds. Also, presenting contradictory information side-by-side is not representative, as these posts would be interspersed with other posts. Our results are indicative of a best-case situation where users carefully consider all relevant information, which is

---

[7]Three for participants assigned Facebook because they were not shown a user handle.

likely closer to the truth in emergency situations when finding good information is safety-critical and social media is saturated with posts about an ongoing event.

For open-response questions, we give the percentage of participants who stated each theme. However, not mentioning a theme does not indicate disagreement. Participants may have failed to state an idea or considered other thoughts more relevant. Our open-response results should be viewed as a measure of what was "front of mind" when answering.

We expect non-U.S. populations' views of verified indicators differ due to the ways social media is politicized in the U.S. [11]. Cross-cultural comparisons require a sample size infeasibly large for this study. Instead, we limit our sample and conclusions to a single culture with which we are familiar.

Even though we used Prolific's census-representative sample feature, Prolific users are often more knowledgeable regarding privacy and security and more likely to use multiple social media platforms [71], which may impact generalizability. To account for these differences, we controlled for social media use and security attitudes in our regressions.

As these limitations apply across all conditions, we focus primarily on between-condition comparisons.

5   Survey Results

The majority of our key findings are taken from our regression analyses over initial perceived correctness (Table 4a), perceived correctness after proposing a new policy (Table 4b), and perceived policy acceptability (Table 5). Only variables in the final selected model are shown (as groups of rows). We give the base case first for categorical variables. We selected base cases expected to correlate with the lowest levels of VA perceived correctness and policy acceptability.

For categorical variables, OR is the odds ratio of the outcome (e.g., acceptability) increasing one Likert-scale unit when switching from the base case to the given parameter level. For numeric variables (e.g., SA-6), OR is the odds the outcome increases one Likert-scale unit for each one-point increase in the numeric variable. For example, the OR for *Police* in Table 4a indicates a participant assigned *Police* instead of *Coffee*—holding all other variables equal—would be $1.57\times$ as likely to increase one unit in perception that the VA posted the correct message. Because this effect is greater than one, participants are more likely to report the VA as correct for *Police* than *Coffee*. *Police*'s confidence interval (CI) indicates that if we ran the study many times, we would expect 95% of runs to produce ORs between 1.31 and 1.87. The *p*-value (< 0.001) is less than our significance threshold ($\alpha = 0.05$), indicating a significant difference between *Police* and *Coffee*.

5.1   Participants

1739 participants attempted and 1660 completed the survey. We removed 27 who failed the attention check, 30 who gave

Metric	%	Metric	%
Age		**Education**	
18-29 years	23.8%	H.S. or below	13.0%
30-49 years	34.9%	Some college/	32.9%
50-64 years	28.9%	Assoc.	
65+ years	12.4%	B.S. or above	53.9%
		Prefer not to respond	0.3%
Platform w/Account			
Facebook	82.2%	**Social Media Use**	
YouTube	78.9%	<30 mins daily	19.1%
Instagram	68.7%	30 mins-1 hr daily	30.6%
Twitter	66.7%	1-2 hrs daily	28.7%
LinkedIn	42.7%	2-4 hrs daily	16.6%
TikTok	37.0%	5-6 hrs daily	3.3%
		>6 hrs daily	1.6%

Table 3: Participant demographics. Percentages may not add to 100% due to non-response or selection of multiple options.

nonsensical or obviously AI-generated responses (long paragraphs with distinctive wording) to open-ended questions, and 3 who withdrew after the debrief. Our final dataset contains 1600 responses (50+ per condition).

Table 3 summarizes participant demographics. Additional demographics are reported in Appendix D. Our participants' gender and income were similar to the 2020 U.S. Census [1]. Participant ethnicities were similar to the U.S. Census, though White participants were overrepresented and Latino/a participants were underrepresented. Participants were more educated and younger on average than the U.S. population, though similar to estimated Twitter user demographics [80]. Participants' average SA-6 score was 3.61, close to the average score from a U.S. Census-representative sample [18].

Participants most often had accounts with Facebook (82.2%), YouTube (78.9%), Instagram (68.7%), and Twitter (66.7%)—similar to other social media use surveys [5]. They most often used Twitter at least every other day (38.8%), with the majority using it at least once per week (64.9%), and many having no account (35.1%). Participants were more active on Facebook, with most using it at least every other day (56.1%) and only 19.8% not having an account. Facebook use did not vary significantly between participants assigned to the Twitter and Facebook conditions ($\chi^2 = 2.9, p = 0.566$). Twitter usage did vary between platform conditions ($\chi^2 = 9.6, p = 0.047$), but the effect size indicates little if any association ($\phi = 0.08$) [13, pg. 282].

5.2   Initial Impact of Verified Account (RQ2)

Here, we discuss participant perceptions of the contradictory posts' credibility (Part B) and how they perceived the verified indicator impacting their decision-making (Part C) *prior* to being given a verification policy. Figure 3 summarizes initial credibility perceptions divided by experimental condition, and Figure 6 in Appendix D summarizes participants' perceptions of the account features' decision-making impact.

No difference between platforms. Across conditions, par-

Variable	Value	Odds Ratio	CI	p-value
Content	Coffee	–	–	–
	Police	**1.56**	**[1.31, 1.87]**	**<0.001***
Position	Contradict.	–	–	–
	Declar.	**1.42**	**[1.19, 1.69]**	**<0.001***
Age	–	–	–	–
	+1	**0.99**	**[0.98, 0.99]**	**<0.001***

– Base case (OR=1, by definition)
*Significant effect

(a) Initial Perceived Verified Account Correctness

Variable	Value	Odds Ratio	CI	p-value
Content	Coffee	–	–	–
	Police	**4.13**	**[3.39, 5.02]**	**<0.001***
Availability	Open	–	–	–
	Notable	**1.80**	**[1.50, 2.17]**	**<0.001***
Verification Method	Phone	–	–	–
	Gov ID	**1.30**	**[1.08, 1.56]**	**0.005***
Facebook User	False	–	–	–
	True	**1.54**	**[1.21, 1.95]**	**<0.001***
SA-6	–	–	–	–
	+1	**1.21**	**[1.08, 1.36]**	**<0.001***

– Base case (OR=1, by definition)
*Significant effect

(b) Verified Account Correctness After Policy Given

Table 4: Summary of regression over participants' VA correctness perception (a) before and (b) after being shown a specific policy. Pseudo R^2 measures for (a) were 0.01 (McFadden) and 0.04 (Nagelkerke), and for (b) were 0.07 (McFadden) and 0.17 (Nagelkerke).

ticipant perceptions of the more likely credible post were evenly distributed. Participants most often indicated the VA was "Definitely" or "Probably" credible (43.9%). However, 32.1% indicated "Either the verified or not VA" was credible and 24.1% chose "Definitely" or "Probably" the non-VA. Results were similar whether participants were assigned Twitter (43.8% VA, 33.3% either, 22.9% non-VA), or Facebook (44.0% VA, 30.8% either, 25.3% non-VA). The selected regression (Table 4a) did not include platform, indicating no observed statistically significant difference between platforms.

When asking participants directly about the verified indicator's impact on their decision-making, responses again were split. A slight majority indicted it had no impact (52.0%), while 48.0% reported at least a "Minor effect." Participants were statistically significantly more likely to rank the verified indicator's effect higher than the account picture ($Z = 14.46, p < 0.001$) and handle ($Z = 7.31, p < 0.001$) according to Wilcoxon-Pratt signed rank tests. We did not observe a statistically significant difference between the verified indicator's and account name's perceived impact ($Z = 1.71, p = 0.087$).

Comparing platforms (Figure 6) there is no clear difference: 46.6% of Facebook-assigned participants reported at least a "Minor effect" versus 49.4% for Twitter. No statistically
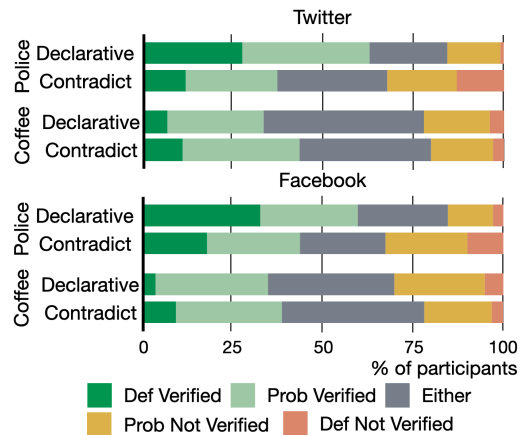
Figure 3: Likert-scale response showing whether participants perceived the VA as more likely credible, organized by assigned social media platform, content type, and the position taken by the VA.
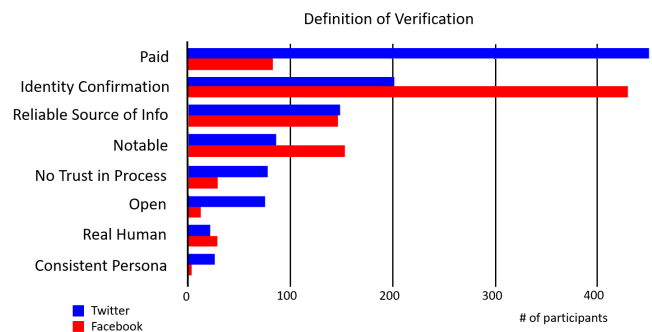
Figure 4: Participants' verification definitions by platform.

significant difference was observed ($\chi^2 = 4.82, p = 0.186$).

Content had the biggest effect. Participants shown the Police content were statistically significantly more likely to perceive the VA as credible ($OR = 1.56, p < 0.001$). If the VA posted the declarative statement (e.g., there *was* a bomb), participants were statistically significantly more likely to perceive the VA as credible ($OR = 1.42, p < 0.001$). This follows prior work [75], which showed content drives message credibility.

Age has some effect. Grouping participants by decade, we observed a downward trajectory in percentage of participants perceiving the VA as "Definitely" or "Probably" credible (52.1% of <30s to 24.1% of >70s). Older participants more often indicated "Either the verified or not VA" was credible (25.8% of <30s to 42.2% of >70s)—the correct response, as VAs do not necessarily post credible content. With each additional year, participants were $0.99\times$ as likely to find the VA more credible by one point ($p < 0.001$). When comparing an individual one standard deviation older ($\sim$15.75 years), we would expect them to be $0.85\times$ as likely to increase one point on the Likert scale. This contradicts Xiao et al.'s prior observation of no statistically significant relationship [81].

5.3 Verification Policy Definitions (RQ2)

Here, we discuss participants' free-response verification definitions (Part D) prior to priming about a particular policy. These definitions mostly aligned with those found via our policy review (Section 3.2). Because we asked participants about platforms with divergent policies (i.e., Facebook and Twitter), we discuss each separately. Our final codebook is in supplementary materials [2]. Figure 4 summarizes responses by platform. Because participants could describe multiple dimensions, these counts do not sum to the total number of participants. These numbers represent front-of-mind definitions; not mentioning a dimension does not necessarily mean the participant does not believe it applies to the policy.

Participants were more likely to believe Facebook confirms user identity. 54.2% of Facebook-assigned participants stated Facebook confirms the user's identity matches their online persona. As one participant said, "[users] need to submit identification, and Facebook manually reviews it." Only 25.1% of Twitter-assigned participants said the same. This difference was statistically significant ($\chi^2 = 140.58, p < 0.001$). While the share of Twitter-assigned participants who believe Twitter verifies identity is concerning, the majority of participants' perceptions align with each platform's actual policies. While not directly comparable, we note the percentage of participants stating Twitter verifies user identity in our survey is much lower than in Xiao et al.'s [81], potentially indicating user understanding of Twitter's policy has improved.

Many participants focused on measures to ensure accounts were made by real humans, not bots. Instead of ensuring VAs' true identity matched their persona, many perceived verification as simply requiring the user verify personal information (e.g., mailing address, email, phone number), limiting verification of bots (18.4% Twitter; 18.3% Facebook). We did not observe a statistically significant difference ($\chi^2 < 0.001, p = 1$). Both platforms require these checks, though they are Twitter's primary verification mechanism.

Payment is mostly associated with Twitter. More than half of Twitter-assigned participants mentioned payment (56.0%). One participant explained, "You pay $8 and elon gives you the blue checkmark." Conversely, few (10.4%) Facebook-assigned participants believed payment was required. This difference was statistically significant ($\chi^2 = 370.6, p < .001$). Xiao et al. found similar results (i.e., Twitter is paid and Facebook is free) [81], but this was correct at the time, as Facebook switched to a paid model after their survey. We show this perception of Facebook as free is now a misconception, indicating users are unaware of Facebook's policy change.

Facebook-assigned participants were more likely to believe verification was for "notable" accounts. 19.3% of Facebook-assigned participants said only notable accounts could be verified. As one participant said, "they have to be notable enough to where other people want to make fake accounts of them." This misconception was not common, but was more common ($\chi^2 = 21.881, p < .001$) among Facebook-assigned than Twitter-assigned participants (10.8%). Conversely, Twitter-assigned participants (8.9%) were more likely ($\chi^2 = 44.80, p < 0.001$) than Facebook-assigned participants (1.4%) to say anyone could be verified.

Facebook-assigned participants were more likely to be unaware of the platform's policy. Many Facebook participants reported not knowing Facebook's policy (17.2%). One participant said, "I actually don't know what the qualifications are to maintain a checkmark. I kind of blindly trust it has been adequately verified." Some were even unaware Facebook had VAs (2.1%). One participant stated, "Facebook uses blue checkmarks? I thought you were talking about Twitter." Many fewer Twitter-assigned participants (7.6%) reported lacking knowledge ($\chi^2 = 32.988, p < .001$).

Some people still conflate verification with credibility. Though not many, some participants (3.7% of Facebook-assigned; 2.7% of Twitter-assigned) continue to believe verification indicates the account is a reliable source of information. As one participant explained, "I would think that Facebook's fact checkers would verify the post was legit and gave good information." This mirrors previous work showing a minority of users conflate authenticity with credibility [14, 22, 75, 81].

Participants criticized Twitter more. Some participants mistrusted the verification process, describing it as politically biased ("They must share the same 'opinion' as Facebook's creator/staff"), failing to prevent inauthentic accounts ("there are so many loopholes now for bots to act like humans and falsify information"), or expressed nihilism ("Better to let the [expletive] thing die than waste time on this verification nonsense"). Criticism was more common ($\chi^2 = 23.914, p < .001$) among Twitter-assigned (9.6%) than Facebook-assigned participants (3.4%). These are small fractions, but we note we asked for participants' definition of the process, not their opinion of it.

5.4 Verification Policy Perceptions (RQ3)

We next discuss perceptions of VA posts' credibility after defining a verification policy (Part F) and how acceptable participants consider the policy (Part E). We saw a significant increase in perceptions that the VA's posted content was credible ($Z = 21.69, p < 0.001$ in Wilcoxon Signed Rank test). This was likely affected by our priming participants to focus on verification by asking for a definition (Part D) and giving a specific policy (Part E). Therefore, we do not compare initial and after-priming responses, but only provide between-participant comparisons on the after-priming question.

We focus first on the three varied policy dimensions (Eligibility, Verification Method, and Payment), then discuss other factors. Figure 5a summarizes participant correctness per-

(a) Post-priming Credibility Perceptions

(b) Policy Acceptability
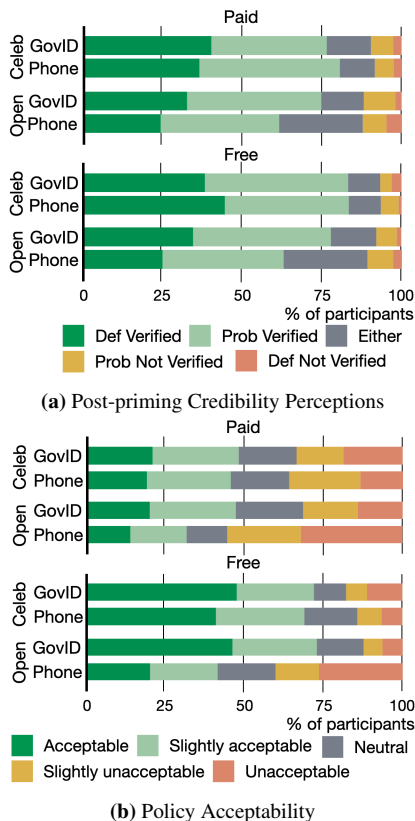
Figure 5: Likert-scale response indicating (a) posted content credibility perceptions and (b) policy acceptability after defining a policy. Both are organized by assigned policy dimensions.

Variable	Value	Odds Ratio	CI	*p*-value
Eligibility	Anyone	–	–	–
	Notable	**1.57**	**[1.32, 1.88]**	**<0.001***
Verification Method	Phone	–	–	–
	Gov ID	**1.80**	**[1.51, 2.15]**	**<0.001***
Payment	Paid	–	–	–
	Free	**2.53**	**[2.11, 3.03]**	**<0.001***
SA-6	1	–	–	–
	+1	**1.27**	**[1.14, 1.41]**	**<0.001***
*Significant effect		– Base case (OR=1, by definition)		

Table 5: Policy acceptability regression summary. The model's pseudo R^2 measures were 0.04 (McFadden) and 0.11 (Nagelkerke).

ceptions, divided by dimension, and Figure 5b shows how acceptable participants considered each policy.

Limiting verification to notable accounts and authenticating with a government ID (govID) increases perceived posted content credibility and acceptability. Most govID-assigned participants (78.0%) believed the VA's posted content was "Definitely" or "Probably" more credible. VA posted content credibility perceptions dropped to 72.0% when told accounts were verified via email or phone, with more participants saying "Either" posted content could be credible (18.5%; 12.9% for govID). This difference was statistically significant, with govID-assigned participants 1.30× more likely to increase one point toward the VA (*p* = 0.005, Table 4b). GovID-assigned participants were also statistically significantly more likely to find the policy acceptable (*OR* = 1.78, *p* < 0.001, Table 5) with a majority finding it "Slightly acceptable" or "Acceptable" (60.0%), but this was a minority opinion for those shown the email or phone policy (46.9%). Requiring govID was the most commonly desired policy change (N=306) with only 33 participants saying govID should not be required. One participant explained, "I would require a photo ID. I can say I'm John Travolta and I can give you my email address (which can be almost any-

thing) to confirm me, but I'm not John Travolta." This aligns with best practices for verification [24], as it is easier to create a new email or phone number than falsify a government document, and there have already been many cases of malicious accounts defeating phone verification [50, 66, 70].

There was a similar difference when comparing notable-only-assigned participants (80.8% "Definitely" or "Probably" more credible), as opposed to participants assigned an open policy (69.3% "Definitely" or "Probably" more credible). This difference was statistically significant with a slightly larger effect size (*OR* = 1.80, *p* < 0.001). Participants reported higher acceptability for the notable-only policy (58.5% "Slightly acceptable" or "Acceptable"), compared to an open policy (48.4% "Slightly acceptable" or "Acceptable")—also statistically significant (*OR* = 1.56, *p* < 0.001). However, when asked for a desired policy change, a greater proportion of participants wanted the policy to be open, not notable. Of the 804 participants shown a notable-only policy, 18.9% wanted it to be open, while only 8.1% of open policy participants wanted verification for notable users only. This sentiment for open policies was driven by concerns of equality; as one participant stated, "I don't believe one has to be well known or high-profile to be verified. That absolutely stinks of elitism." This contradicts our regression results, suggesting participants are split on their preference for Eligibility.

Payment does not affect perceived correctness, but reduces approval. We did not observe a statistically significant impact on participants' VA posted content credibility perceptions based on payment. When shown a free verification policy, 76.8% of participants indicated the VA's post was "Definitely" or "Probably" more credible, compared to 73.2% of participants shown a paid policy. Free verification was the strongest factor increasing policy approval (*OR* = 2.54, *p* < 0.001). While 64.0% shown a free policy found it at least "Slightly acceptable", only 43.1% said the same of paid policies. Like Xiao et al. [81], we found many participants focused on price when suggesting a policy change (N=342). One participant said, "Money shouldn't be a barrier to doing public good." This indicates payment might not impact users' VA percep-

tions, but it displeases users, as observed with Twitter [30].

Social media use and security attitudes play a role. Participants who use Facebook were more likely to view the VA's posted content as credible (76.9% said "Definitely" or "Probably" more credible) compared to non-Facebook users (66.3% said "Definitely" or "Probably" more credible) ($OR = 1.54, p < 0.001$). Participants who reported taking more general security actions were more likely to view the VA's posted content as more credible ($OR = 1.21, p < 0.001$) and find the policy acceptable ($OR = 1.27, p < 0.001$). This may suggest the misconception that VAs are "secure," i.e., should be trusted over other accounts. However, prior work contradicts this [22, 75, 81], and few participants said verification indicates credibility (see Section 5.3). This may instead be an effect of the specific contrasting scenarios we chose, where the only major difference was the verified indicator and accounts were authoritative. Security-conscious participants may have been more likely to consider this difference.

6 Discussion

Our results reveal users' understanding of recent verification policy changes, along with their perceptions of the changes and other potential policies. We suggest social media platform verification policy improvements and discuss future work.

Many participants were aware of Twitter's transition to paid, open verification without a required identity check. While the results are not directly comparable, this seems to indicate improved user awareness relative to Xiao et al.'s earlier survey, which found many users believed Twitter performed rigorous identity checks [81]. Conversely, our participants were unaware of Facebook's policy changes, believing it remained free and restricted to notable accounts. This misunderstanding is not as consequential as incorrectly believing accounts undergo identity verification. However, our results suggest participants were more likely to perceive VA posts as credible when only notable accounts are verified, so this misunderstanding still introduces misplaced trust.

To improve trust in the verification process, platforms should employ rigorous ID checks. Participants were more likely to find the VA's posted content more credible when it was verified with a govID, more likely to find govID verification acceptable, and frequently suggested an ID check be added to improve verification. This shows users value identity verification over other requirements for bot prevention or account consistency. If Twitter reverts to rigorous identity checks (as was rumored [53]), future work should consider whether perceptions of Twitter's policy improve, as our hypothetical settings suggest, or if these perceptions represent a one-way-ratchet and are already ingrained in users' minds.

We did not observe any significant difference in the verified indicator's effect between platforms before priming about verification. When primed, participants shown Facebook's policy were statistically significantly more likely to find the VA more credible. This suggests users do not consider policy differences without priming, and because Facebook's policy is less well known, may default to their understanding from Twitter. As social media platforms change verification policies, they must educate users to avoid misunderstandings. This is especially true when changing govID and notability requirements, which had significant impacts, though future work must determine how best to educate users.

Restrictions on account eligibility produced mixed results. Under a notable-only policy, participants were more likely to perceive the VA's posted content as more credible and find the policy acceptable. However, when asked to suggest changes to the platform, participants contradicted this sentiment by saying verification should be open to all users. One remedy suggested by a few participants (N=19) is a tiered approach to verification. As one participant suggested, "I think for public service accounts such as the fire department, police department, federal government, etc. there should be a more rigorous verification process." Similarly, some participants wanted the platform to evaluate users' authoritative credentials (N=62). This could include verifying hospital credentials of medical professionals or press credentials for journalists. Twitter somewhat employs this approach with special indicators for government and business accounts (✓, ✓). Although users may prefer this in theory, prior work found users misunderstood both badges [81]. Future research should consider the impact of these indicators, especially in emergency situations when an account's authority is important (similar to our bomb threat examples) and under various Verification Method regimes to determine the interaction between these variables.

Perhaps the most polarizing verification change is switching to a paid model. Participants found paid policies unacceptable and wanted to remove payment, matching prior work [81]. However, we did not observe an effect from payment on participants' posted content credibility perceptions. We might have expected participants to be less likely to trust paying accounts, since Twitter's verified indicator has been described as a "scarlet letter" [30] and impostor accounts have been created [50]. However, it seems users correctly associate these problems with the lack of identity verification, not payment. This suggests that while payment might annoy users, it does not negatively impact how they evaluate VA posts.

Finally, participants were statistically significantly more likely to find the VA credible after priming about verification. This could be the result of asking participants to consider a hypothetical policy, but appears more likely due to priming effects. This could be problematic for platforms using policies that do not have rigorous identity verification. Malicious users may be able to fool others into believing their posts by drawing attention to their verified indicator. Future work should investigate situations where other information beyond the verified indicator varies between contradictory posts to measure the potential risk of social engineering attacks.

Acknowledgment

We thank the anonymous reviewers who provided very helpful comments on drafts of this paper.

References

[1] Census bureau data. https://data.census.gov/. (Accessed 08-30-2023).

[2] Social media account verification perceptions. https://doi.org/10.17605/OSF.IO/A9Y3J.

[3] Internet Archive. Internet archive: Wayback machine. https://archive.org/web/. (Accessed 09-10-2023).

[4] Associated Press. Students criticize the University of North Carolina's response to an active shooter emergency. https://www.voanews.com/a/awash-in-social-media-how-are-us-police-learning-to-inform-the-public-better-after-shootings-/7100938.html, 2023. (Accessed 09-10-2023).

[5] Brooke Auxier and Monica Anderson. Social media use in 2021. https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/, 2021. (Accessed 08-01-2019).

[6] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

[7] Cody Buntain, Jennifer Golbeck, Brooke Liu, and Gary LaFree. Evaluating public response to the boston marathon bombing and other acts of terrorism through twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1):555–558, Aug. 2021.

[8] Pew Research Center. Social media and news fact sheet. https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/, 2022. (Accessed 08-18-2023).

[9] Shelly Chaiken. The Heuristic Model of Persuasion. In *Social Influence: the Ontario Symposium*, volume 5, pages 3–39, 1987.

[10] Brian X. Chen and Ryan Mac. Twitter's blue check apocalypse is upon us. here's what to know. The New York Times. https://www.nytimes.com/2023/03/31/technology/personaltech/twitter-blue-check-musk.html, 2023. (Accessed 09-10-2023).

[11] Laura Clancy. Americans differ by party, ideology over the impact of social media on U.S. democracy, December 2022.

[12] Juliet Corbin, Anselm Strauss, and Anselm L Strauss. *Basics of qualitative research*. Sage, 2014.

[13] H. Cramér. *Mathematical Methods of Statistics*. Princeton Landmarks in Mathematics and Physics. Princeton University Press, 1999.

[14] Jazlyn Elizabeth Dumas and Rusty Allen Stough. When influencers are not very influential: The negative effects of social media verification. *Journal of Consumer Behaviour*, 21(3):614–624, 2022.

[15] @elonmusk. Twitter's current lords & peasants system for who has or doesn't have a blue checkmark is bullshit. Power to the people! Blue for $8/month. https://twitter.com/elonmusk/status/1587498907336118274, 2022. (Accessed 09-10-2023).

[16] @elonmusk. Yes, this will destroy the bots. If a paid Blue account engages in spam/scam, that account will be suspended. . . . . https://twitter.com/elonmusk/status/1587512669359292419, 2022. (Accessed 09-10-2023).

[17] @elonmusk. Given that modern AI can solve any "prove you're not a robot" tests, it's now trivial to spin up 100k human-like bots. . . . https://twitter.com/elonmusk/status/1640199090112806912?s=20, 2023. (Accessed 09-10-2023).

[18] Cori Faklaris, Laura A. Dabbish, and Jason I. Hong. A Self-Report measure of End-User security attitudes (SA-6). In *Fifteenth Symposium on Usable Privacy and Security*, SOUPS 2019, pages 61–77, Santa Clara, CA, August 2019. USENIX Association.

[19] Andrew J. Flanagin and Miriam J. Metzger. The Role of Site Features, User Attributes, and Information Verification Behaviors on the Perceived Credibility of Web-Based Information. *New Media & Society*, 9(2):319–342, 2007.

[20] B. J. Fogg, Cathy Soohoo, David R. Danielson, Leslie Marable, Julianne Stanford, and Ellen R. Tauber. How do users evaluate the credibility of web sites? a study with over 2,500 participants. In *Proceedings of the 2003 Conference on Designing for User Experiences*, DUX '03, page 1–15, New York, NY, USA, 2003. Association for Computing Machinery.

[21] Brian Fung. How Elon Musk transformed Twitter's blue check from status symbol into a badge of shame. CNN Business. https://www.cnn.com/2023/04/24/tech/musk-twitter-blue-check-mark/index.html, 2023. (Accessed 09-10-2023).

[22] Christine Geeng, Savanna Yee, and Franziska Roesner. Fake news on facebook and twitter: Investigating how people (don't) investigate. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA, 2020. Association for Computing Machinery.

[23] Google. Verification Badges on Channels. https://support.google.com/youtube/answer/3046484. (Accessed 09-10-2023).

[24] Paul A. Grassi, James L. Fenton, Naomi B. Lefkovitz, Jamie M. Danker, Yee-Yin Choong, Kristen K. Greene, and Mary F. Theofanos. NIST Special Publication 800-63A, Digital Identity Guidelines, Enrollment and Identity Proofing. Technical report, National Institute of Standards and Technology, 06 2017.

[25] Amelia Hassoun, Ian Beacock, Sunny Consolvo, Beth Goldberg, Patrick Gage Kelley, and Daniel M. Russell. Practicing information sensibility: How gen z engages with online information. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery.

[26] Andrew F. Hayes and Klaus Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89, 2007.

[27] Brian Hilligoss and Soo Young Rieh. Developing a Unifying Framework of Credibility Assessment: Construct, Heuristics, and Interaction in Context. *Information Processing & Management*, 44(4):1467–1484, 2008.

[28] Kyle Hunt, Bairong Wang, and Jun Zhuang. Misinformation debunking and cross-platform information sharing through twitter during hurricanes harvey and irma: a case study on shelters and id checks. *Natural Hazards*, 103:861–883, 2020.

[29] Christian Johnson and William Marcellino. Reining in COVID-19 Disinformation from China, Russia, and Elsewhere, November 2021. https://www.rand.org/blog/2021/11/reining-in-covid-19-disinformation-from-china-russia.html.

[30] Alex Kirshner. How Elon Musk Turned the Blue Check Mark Into a Scarlet Letter. Slate. https://slate.com/technology/2023/04/elon-musk-twitter-blue-check-marks-verification-lebron-james.html, 2023. (Accessed 09-10-2023).

[31] Ziva Kunda. The Case for Motivated Reasoning. *Psychological bulletin*, 108(3):480, 1990.

[32] Ian Lamont. Plane Lands on the Hudson, and Twitter Documents It All. Computerworld. https://www.computerworld.com/article/2530453/plane-lands-on-the-hudson--and-twitter-documents-it-all.html, 2009. (Accessed 09-10-2023).

[33] Annabelle Liang. Elon Musk: Twitter boss announces blue tick shake-up. BBC News. https://www.bbc.com/news/business-65095684, 2023. (Accessed 09-10-2023).

[34] LinkedIn. Verified on your linkedin profile. https://www.linkedin.com/help/linkedin/answer/a1359065. (Accessed 09-10-2023).

[35] Megan Loe. 5 VERIFIED Ways You Can Fact-check Online Claims. Verify. https://www.verifythis.com/article/news/verify/fact-sheets-verify/5-tips-fact-check-online-claims-yourself-guide/536-64c58fc6-f17d-42dd-9970-b1b4814f9a87, 2023. (Accessed 09-06-2023).

[36] Gary Machado, Alexandre Alaphilippe, Roman Adamczyk, and Antoine Gregoire. Indian Chronicles: deep dive into a 15-year operation targeting the EU and UN to serve Indian interests. Technical report, EU Disinfo Lab.

[37] Odanga Madung and Brian Obilo. Inside the Shadowy World of Disinformation-for-hire in Kenya. Technical report, Mozilla Foundation. Section: Fellowships & Awards.

[38] Miriam Matthews, Katya Migacheva, and Ryan Andrew Brown. Superspreaders of Malign and Subversive Information on COVID-19: Russian and Chinese Efforts Targeting the United States. Technical report, RAND Corporation, April 2021.

[39] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23, nov 2019.

[40] Adam W. Meade and S. Bartholomew Craig. Identifying careless responses in survey data. *Psychological methods*, 17(3):437, 2012.

[41] Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. Twitter under Crisis: Can We Trust What We RT? In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, page 71–79, New York, NY, USA, 2010. Association for Computing Machinery.

[42] Meta. Meta verified | get a verified blue check on instagram, facebook. https://about.meta.com/technologies/meta-verified/. (Accessed 09-10-2023).

[43] Miriam J. Metzger, Andrew J. Flanagin, and Ryan B. Medders. Social and Heuristic Approaches to Credibility Evaluation Online. *Journal of Communication*, 60(3):413–439, 08 2010.

[44] Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, and Julia Schwarz. Tweeting is believing?: Understanding microblog credibility perceptions. In *Proceedings of the 15th ACM Conference on Computer Supported Cooperative Work*, CSCW '12, pages 441–450, New York, NY, USA, 2012. ACM.

[45] Casey Newton and Zoe Schiffer. Elon Musk ignored Twitter's internal warnings about his paid verification scheme. The Verge. https://www.theverge.com/2022/11/14/23459244/twitter-elon-musk-blue-verification-internal-warnings-ignored, 2022. (Accessed 09-10-2023).

[46] Matt Novak. Viral Video Alleging Canadian Wildfires Were 'Set Up' Is Very Misleading. Forbes. https://www.forbes.com/sites/mattnovak/2023/06/09/viral-video-alleging-canadian-wildfires-were-set-up-is-very-misleading/?sh=67e194bb7350, 2023. (Accessed 09-10-2023).

[47] Matt O'Brien. Canada wildfire evacuees can't get news media on facebook and instagram. some find workarounds. AP News. https://apnews.com/article/canada-wildfires-yellowknife-nwt-facebook-instagram-meta-723687efe632884e4eb1172528abb43f, 2023. (Accessed 09-10-2023).

[48] Matt O'Brien and Kathleen Foody. Confusion as Musk's Twitter yanks blue checks from agencies. AP News. https://apnews.com/article/twitter-elon-musk-blue-checkmark-celebrities-544cfd66ed3a62f51a8a80c20e11ac5b, 2023. (Accessed 09-10-2023).

[49] Kari Paul. Russian disinformation surged on social media after invasion of Ukraine, Meta reports. *The Guardian*, April 2022. https://www.theguardian.com/world/2022/apr/07/propaganda-social-media-surge-invasion-ukraine-meta-reports.

[50] Kari Paul. Fake accounts, chaos and few sign-ups: the first day of twitter blue was messy. The Guardian. https://www.theguardian.com/technology/2023/apr/21/elon-musk-twitter-blue-rollout, 2023. (Accessed 09-10-2023).

[51] Karl Pearson. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50(302):157–175, 1900.

[52] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70:153 – 163, 2017.

[53] Sarah Perez. Twitter testing government ID-based verification, new screenshots show. TechCrunch. https://techcrunch.com/2023/03/20/twitter-testing-government-id-based-verification-new-screenshots-show/, 2023. (Accessed 08-18-2023).

[54] Karena Phan. Social media videos push baseless conspiracy theory that blue items were spared from maui wildfires. AP News. https://apnews.com/article/fact-check-conspiracy-blue-items-maui-wildfires-118319149774, 2023. (Accessed 09-10-2023).

[55] Monica Prasad, Andrew J. Perrin, Kieran Bezila, Steve G. Hoffman, Kate Kindleberger, Kim Manturuk, and Ashleigh Smith Powers. "there must be a reason": Osama, saddam, and inferred justification. *Sociological Inquiry*, 79(2):142–162, 2009.

[56] Prolific. Representative samples. https://researcher-help.prolific.co/hc/en-gb/articles/360019236753-Representative-samples, 2023. (Accessed 09-10-2023).

[57] Adrian E. Raftery. Bayesian model selection in social research. *Sociological methodology*, pages 111–163, 1995.

[58] Harry T. Reis and Charles M. Judd. *Handbook of research methods in social and personality psychology*. Cambridge University Press, 2000.

[59] David E. Sanger and Julian E. Barnes. U.S. Warns Russia, China and Iran Are Trying to Interfere in the Election. Democrats Say It's Far Worse. *The New York Times*, July 2020. https://www.nytimes.com/2020/07/24/us/politics/election-interference-russia-china-iran.html.

[60] Craig Silverman. Verification and Fact Checking. European Journalism Centre. https://datajournalism.com/read/handbook/verification-1/additional-materials/verification-and-fact-checking. (Accessed 09-06-2023).

[61] Craig Silverman and Rina Tsubaki. A guide to verifying digital content in emergencies. Global Investigative Journalism Network. https://gijn.org/2014/03/18/a-guide-to-verifying-digital-content-for-emergency-coverage/, 2014. (Accessed 09-06-2023).

[62] Snapchat. How to verify your public profile. https://businesshelp.snapchat.com/s/article/public-profile-verify?language=en_US. (Accessed 09-10-2023).

[63] Elliott Sober. Instrumentalism, parsimony, and the akaike framework. *Philosophy of Science*, 69(S3):S112–S123, 2002.

[64] Truth Social. Red check verification. https://help.truthsocial.com/moderation/how-to-get-verified. (Accessed 09-10-2023).

[65] Todd Spangler. Elon Musk Says Twitter 'Final Date' for Removing Legacy Blue Check-Marks Is 4/20. Variety. https://variety.com/2023/digital/news/twitter-musk-date-removal-blue-checkmarks-legacy-1235570782/, 2023. (Accessed 09-10-2023).

[66] Mariana Spring and Laura Gozzi. Twitter blue tick: Multiple hillarys and new yorks as verifications disappear. BBC News. https://www.bbc.com/news/technology-65346263, 2023. (Accessed 09-10-2023).

[67] Biz Stone. Not Playing Ball. Twitter. https://blog.twitter.com/official/en_us/a/2009/not-playing-ball.html, June 2009. (Accessed 07-18-2023).

[68] Anselm Strauss and Juliet Corbin. *Basics of qualitative research*, volume 15. Newbury Park, CA: Sage, 1990.

[69] S. Shyam Sundar. The MAIN Model: A Heuristic Approach to Understanding Technology Effects on Credibility. *Digital media, youth, and credibility*, 73100, 2008.

[70] Pete Syme. Elon musk's war against twitter bots isn't going very well. next, you'll have to pay to dm those who don't follow you. Business Insider. https://www.businessinsider.com/elon-musk-war-on-twitter-bots-isnt-working-limits-dms-2023-6, 2023. (Accessed 09-10-2023).

[71] Jenny Tang, Eleanor Birrell, and Ada Lerner. Replication: How well do my results generalize now? the external validity of online privacy and security surveys. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 367–385, Boston, MA, August 2022. USENIX Association.

[72] TikTok. Verified accounts on tiktok. https://support.tiktok.com/en/using-tiktok/growing-your-audience/how-to-tell-if-an-account-is-verified-on-tiktok. (Accessed 09-10-2023).

[73] Twitter. How To Get the Blue Checkmark on X. Twitter. https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts. (Accessed 09-10-2023).

[74] Twitter. Legacy verification policy. https://help.twitter.com/en/managing-your-account/legacy-verification-policy. (Accessed 09-10-2023).

[75] Tavish Vaidya, Daniel Votipka, Michelle L. Mazurek, and Micah Sherr. Does being verified make you more credible? account verification's effect on tweet credibility. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, New York, NY, USA, 2019. Association for Computing Machinery.

[76] Sarah Vieweg. Microblogged contributions to the emergency arena: Discovery, interpretation and implications. *Computer Supported Collaborative Work*, pages 515–516, 2010.

[77] James Vincent. Twitter says paying blue subscribers now get 'prioritized rankings in conversations'. https://www.theverge.com/2022/12/23/23523845/twitter-blue-paying-priority-replies-conversations, 2022. (Accessed 09-10-2023).

[78] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.

[79] Sam Wineburg, Sarah McGrew, Joel Breakstone, and Teresa Ortega. Evaluating Information: The Cornerstone of Civic Online Reasoning. *Stanford Digital Repository*, 2016.

[80] Stefan Wojcik and Adam Hughes. Sizing Up Twitter Users. Pew Research Center. https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/, 2019. (Accessed 08-01-2019).

[81] Madelyne Xiao, Mona Wang, Anunay Kulshrestha, and Jonathan Mayer. Account verification on social media: User perceptions and paid enrollment. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 3099–3116, Anaheim, CA, August 2023. USENIX Association.

A Overview

In our appendices, we describe our web scraping process for policy collection (Section B), provide our survey text (Section C), and additional tables and figures not included in the main paper for brevity (Section D. The full set of mock posts shown to users in our survey, the full codebook of free-response questions, demographic questions, debrief text, and the timeline of policy changes we observed can be found at https://osf.io/a9y3j/?view_only=d2608dffe87f40c09885c4e55637ddeb.

B Policy Review Web Scraping Process

To capture each platform's verification policies, we created a simple web scraper in Python using the BeautifulSoup4 and Selenium libraries. This script was run daily to pull each policy, compare it to the prior version, and record changes. Because we began our

collection in February 2023, we used the Internet Archive's Wayback Machine [3] to collect older changes to the platforms' policies. Therefore, our review could be an under-approximation of changes in the period prior to our direct collection. However, we note that we were able to capture all major changes to Twitter reported in the news, and no other platform had major changes during this period. This process generated a dataset of timestamped verification policy changes for each platform.

C Survey Questionnaire

In this appendix, we provide the full text of our survey for one particular condition (Twitter post with police content with the verified indicatorassigned to the declarative statement). Throughout, we provide heading indicating the section of the survey as shown in Figure 1. These headings were not included in the survey shown to participants and are only included here for readability.

------------------ *Survey begins* ------------------

(Consent, Part A)
[Survey Consent presented here]

------------------ *page break* ------------------

In this study, we will display a pair of social media posts and ask you questions about the content shared in the posts.

------------------ *page break* ------------------

(Implicit Measure of Credibility, Part B)

Sherling Police Dept. ✔
@SherlingPolice

ALERT: We are currently investigating an active bomb threat in the Downtown area shopping plaza. Please avoid the area until further notice. All personnel in and around the area are to evacuate immediately.

5:10 PM · 1 May 2023

Post A

Sherling Police Department
@SherlingPD

ALERT: Reports of an active bomb threat in the Downtown area shopping plaza are false. We encourage everyone to go about their day normally, thank you for your cooperation.

5:10 PM · 1 May 2023

Post B

Please answer the following questions considering the two Twitter posts above.

1. Post A and Post B contain conflicting information. Which of the posts do you believe is correct?

   (a) Definitely A

   (b) Probably A

   (c) Equally likely to be A or B

   (d) Probably B

   (e) Definitely B

------------------ *page break* ------------------

In this section, we will ask you some questions about how you determined which Twitter post was more correct in the previous section. Specifically, we will highlight different elements of the post and ask you how much each element influenced your decision. To help you know which visual element we're asking about, we show a different Twitter post, distinct from the posts you saw before, and highlight the element in question.

A VA is denoted by a blue checkmark shown next to the display name, as illustrated within the red box below:

1. On the last page, we asked you which of two contradictory posts was more likely to be correct. When making that choice, how much did the presence of this verified account indicator (✓) affect your decision?

   (a) No Effect

   (b) Minor Effect

   (c) Moderate Effect

   (d) Major Effect

Every post on Twitter includes the display of the user's profile picture next to their handle or username, as exemplified by the red box in the example post below:

1. On the last page, we asked you which of two contradictory posts was more likely to be correct. When making that choice, how much did the account's *profile picture* affect your decision?

   (a) No Effect

   (b) Minor Effect

   (c) Moderate Effect

   (d) Major Effect

A display name is used to identify the account and can differ from the username. On Twitter, it appears next to the account's profile picture as shown by the red box in the example post below:

1. On the last page, we asked you which of two contradictory posts was more likely to be correct. When making that choice, how much did the account's *display name* affect your decision?

   (a) No Effect

   (b) Minor Effect

   (c) Moderate Effect

   (d) Major Effect

On Twitter, a user's handle (also known as their username) is presented next to their profile picture on every tweet they post, and it is marked by the "@" symbol. An example of a user's handle is provided in the red box below:

1. On the last page, we asked you which of two contradictory posts was more likely to be correct. When making that choice, how much did the account's handle affect your decision?

   (a) No Effect

   (b) Minor Effect

   (c) Moderate Effect

   (d) Major Effect

------------------ *page break* ------------------

(Participants' Definition of Verification, Part D)
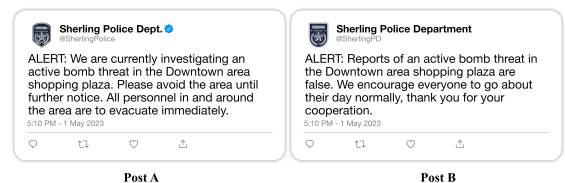
One of the tweets you were previously shown was by an account with a verification checkmark (✓) indicating that the account has been verified.

---

1. Based on your understanding of Twitter's account verification, what requirements must an account satisfy to become verified and obtain a verified checkmark?

------------------ *page break* ------------------

(Show Assigned Policy, Policy Acceptability, Suggested Changes, Part E)

Suppose Twitter adopted a verification policy in which the account had to meet all of the following criteria:

- **Any user** on the platform is allowed to apply for verification Accounts must submit government-issued identification that matches the name of the account being verified

- Any user on the platform is allowed to apply for verification Accounts must **submit government-issued identification** that matches the name of the account being verified

- Accounts **pay a monthly subscription fee** to maintain their verification checkmark

1. To what level do you believe these verification requirements are acceptable for verifying account owner identity?
   (a) Unacceptable
   (b) Slightly Unacceptable
   (c) Neutral
   (d) Slightly Acceptable
   (e) Acceptable

2. If you could suggest one thing to add, remove, or change in this policy to improve its ability in verifying the account owner is who they say they are, what would it be? Please explain why.

------------------ *page break* ------------------

(Credibility Measure After Policy Priming, Part F)

We will now ask you to revisit the Twitter posts you were shown previously, and answer the following questions assuming this new policy was used for verification.

We display the Twitter posts and the new verification policy below for you to reference while you answer the questions.
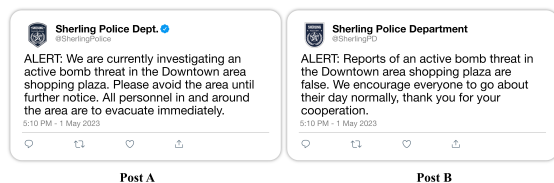
Post A   Post B

- **Any user** on the platform is allowed to apply for verification Accounts must submit government-issued identification that matches the name of the account being verified

- Any user on the platform is allowed to apply for verification Accounts must **submit government-issued identification** that matches the name of the account being verified

- Accounts **pay a monthly subscription fee** to maintain their verification checkmark

1. Which of the following most closely resembles the subject matter of the two posts?
   (a) Police investigating a bomb threat
   (b) Effects of coffee on health
   (c) Food recall due to E. coli outbreak

2. After reviewing the criteria required for an account to receive a verification checkmark, which of the posts do you believe is correct?
   (a) Definitely A
   (b) Probably A
   (c) Equally likely to be A or B
   (d) Probably B
   (e) Definitely B

3. If a friend of yours was unsure about which post to trust, what would you say to this friend to help them decide?

------------------ *page break* ------------------

(Social Media Use, Part G)

Now we will end the survey with several short questions concerning your social media use and and demographics.

1. Which of the following social media platforms do you currently have an account with? Select all that apply.
   - Twitter
   - Facebook
   - Instagram
   - LinkedIn
   - TikTok
   - YouTube
   - Other (please specify)

2. How often do you use Twitter in any given week?
   (a) Daily
   (b) Every other day
   (c) Every two days
   (d) Once a week
   (e) I do not use Twitter

3. How often do you use Facebook in any given week?
   (a) Daily
   (b) Every other day
   (c) Every two days
   (d) Once a week
   (e) I do not use Facebook

4. How much time do you spend on social media sites per day?
   (a) Less than 30 minutes

(b) 30 minutes-1 hour

(c) 1-2 hours

(d) 2-4 hours

(e) 5-6 hours

(f) Greater than 6 hours

------------------ *page break* ------------------

(Security Attitudes, Part H)

Each statement below describes how a person might feel about the use of security measures. Examples of security measures are laptop or tablet passwords, spam email reporting tools, software updates, secure web browsers, fingerprint ID, and anti-virus software.

Please indicate the degree to which you agree or disagree with each statement. In each case, make your choice in terms of how you feel right now, not what you have felt in the past or would like to feel.

1. I seek out opportunities to learn about security measures that are relevant to me

   (a) Strongly disagree

   (b) Somewhat disagree

   (c) Neither disagree nor agree

   (d) Somewhat agree

   (e) Strongly agree

2. I am extremely motivated to take all the steps needed to keep my online data and accounts safe.

   (a) Strongly disagree

   (b) Somewhat disagree

   (c) Neither disagree nor agree

   (d) Somewhat agree

   (e) Strongly agree

3. Generally, I diligently follow a routine for security practices.

   (a) Strongly disagree

   (b) Somewhat disagree

   (c) Neither disagree nor agree

   (d) Somewhat agree

   (e) Strongly agree

4. I often am interested in articles about security threats.

   (a) Strongly disagree

   (b) Somewhat disagree

   (c) Neither disagree nor agree

   (d) Somewhat agree

   (e) Strongly agree

5. I always pay attention to experts' advice about the steps I need to take to keep my online data and accounts safe.

   (a) Strongly disagree

   (b) Somewhat disagree

Factor	Description	Baseline
Posted content Variables		
Platform	The assigned visual design used to display posts	Twitter
Content type	The assigned content condition	Coffee
Position	The side of the argument the verified indicator was assigned to	Contradict.
Policy Variables[1]		
Availability	Who can become verified?	Open
Verification Method	How are accounts verified?	Phone
Payment	Is payment required to become verified?	Paid
Social Media Experience		
Twitter experience	Does the participant report using Twitter (binary)	False
Facebook experience	Does the participant report using Facebook (binary)	False
Social Media Accts.	Number of social media platforms participants use	–
Demographics		
SA-6	Participant's score on Faklaris et al.'s SA-6 scale [18]	–
Age	Age of participant	–
Gender	Gender of participant	Male
Education	Does the participant hold a B.S. or higher degree (binary)	False

[1] Policy variables were only included when considering participants' policy acceptability rating (Part E) and their credibility perceptions after providing them with a mock policy (Part F).

Table 6: Factors used in regression models. Categorical variables are compared individually to the given baseline.

   (c) Neither disagree nor agree

   (d) Somewhat agree

   (e) Strongly agree

6. I am extremely knowledgeable about all the steps needed to keep my online data and accounts safe.

   (a) Strongly disagree

   (b) Somewhat disagree

   (c) Neither disagree nor agree

   (d) Somewhat agree

   (e) Strongly agree

D   Additional Tables and Figures

Finally, we provide tables and figures excluded from the main text for brevity. This includes a summary of the variables in the initial model for each regression (Table 6), additional participant demographics information (Table 7), and a summary of participants' responses regarding perceive impact of each account feature (Figure 6).

Metric	%
Gender	
Woman	49.9%
Man	48.4%
Non-binary	1.2%
Transgender/	0.3%
Agender	
Other	0.2%
Race/Ethnicity	
White	73.9%
Black	11.6%
Asian	6.0%
Hispanic or Latino/a	4.9%
Indigenous	0.7%
Two or More Races	2.0%
Other	0.2%
Prefer not to respond	0.6%

Metric	%
Income	
<$10k	10.6%
$10k-$25k	14.8%
$25k-$50k	25.1%
$50k-$75k	19.1%
$75k-$100k	11.9%
$100k-$150k	10.4%
$150k+	5.1%
Prefer not to respond	3.1%

Table 7: Additional participant demographics.
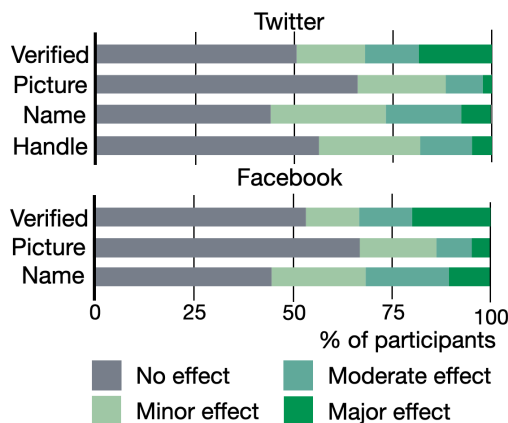
Figure 6: Likert-scale response indicating how much participants perceived each account feature impacted their credibility decision, organized by assigned social media platform.

E   Demographics Questions & Debrief

(Demographics, Part I)

1. What is your age?

2. How do you describe your gender identity?

   (a) Female

   (b) Male

   (c) Agender

   (d) Non-binary

   (e) Gender-queer

   (f) Not sure

   (g) Not listed above [with text entry]

   (h) Prefer not to respond

3. Do you identify as Hispanic and/or Latino?

   (a) Yes

   (b) No

   (c) Prefer not to respond

4. What level of education have you attained?

   (a) Less than high school

   (b) High School graduate (high school diploma or equivalent such as GED)

   (c) Some college, but no degree

   (d) Associate Degree

   (e) Bachelor's Degree

   (f) Master's Degree

   (g) Professional Master's Degree (JD, MD)

   (h) Doctorate Degree

   (i) Prefer not to respond

5. What was your 2020 taxed income?

   (a) Less than $10,000

   (b) $10,000-$24,999

   (c) $25,000-$49,000

   (d) $50,000-$74,999

   (e) $75,000-$99,999

   (f) $100,000-$149,000

   (g) $150,000 and greater

   (h) Prefer not to respond

6. Do you get the majority of your earnings from Prolific or similar platforms?

   (a) Yes

   (b) No

   (c) Prefer not to respond

-------------------- *page break* --------------------

(Debrief, Part J)

Throughout this study you were shown social media posts showing conflicting reports about a particular event or research findings. These events are completely fictional and not based on any true events or findings. For the purpose of this study, these were made up to avoid bias in participant responses.

You were also given a set of criteria used for social media verification. Although the verification criteria we used for this study was based on the verification criteria Twitter and Facebook use to verify accounts on their platforms, the criteria you saw does not reflect the true criteria Twitter and Facebook use for their verification policies.

The verification process Twitter uses can be viewed in full by following this link. In this policy, verification is open to anyone but requires the owner of the account to pay a monthly fee to maintain the verification checkmark. The account must have a display name and profile photo. This display name and profile photo cannot be modified once the account has been verified. The account owner also must confirm a phone number with Twitter. Additionally, the account must show no signs of engaging in platform manipulation or spam, and show no signs of being misleading or deceptive.

The verification process Facebook uses can be viewed in full by following this link. This process is used for verifying accounts owned by public figures, celebrities, or notable brands. Notable brands are those that represent well-known, often searched for brands that are unique (i.e. be the only presence of this business), authentic (i.e. registered business), and have a complete Facebook Page or Facebook Profile (i.e. the account has a completed "About" section, has shared at least one post, and show recent activity.

Facebook also offers account profile verification for all accounts via Meta Verification. To be eligible for Meta Verification the account owner must be at least 18 years of age, have a public or private Facebook profile with the account owner's full name and a profile picture that matches a government issued ID. Additionally, the account must

have a prior posting history, have two-factor authentication enabled. You can learn more about Meta Verification and its process here.

It can be difficult to determine whether information garnered online is true or false. However, there are steps you can take to help confirm if the information you read online is true or meant to mislead you. We provide links to several guides below for verifying digital content and fact checking information online below:

- 5 Ways You Can Fact-Check Online Claims

- A Guide to Verifying Digital Content in Emergencies

- Verification and Fact Checking - A General Guide

"Violation of my body:"
Perceptions of AI-generated non-consensual (intimate) imagery

Natalie Grace Brigham
University of Washington

Miranda Wei
University of Washington

Tadayoshi Kohno
University of Washington

Elissa M. Redmiles
Georgetown University

Abstract

AI technology has enabled the creation of deepfakes: hyper-realistic synthetic media. We surveyed 315 individuals in the U.S. on their views regarding the hypothetical non-consensual creation of deepfakes depicting them, including deepfakes portraying sexual acts. Respondents indicated strong opposition to creating and, even more so, sharing non-consensually created synthetic content, especially if that content depicts a sexual act. However, seeking out such content appeared more acceptable to some respondents. Attitudes around acceptability varied further based on the hypothetical creator's relationship to the participant, the respondent's gender and their attitudes towards sexual consent. This study provides initial insight into public perspectives of a growing threat and highlights the need for further research to inform social norms as well as ongoing policy conversations and technical developments in generative AI.

1 Introduction

Technological advancements in artificial intelligence (AI) have enabled the creation of hyper-realistic synthetic media known as "deepfakes." This term, a portmanteau of "deep learning" and "fake," refers to synthetic image, audio, or video representations of individuals that has been automatically generated using machine learning [31, 49, 89]. Deepfakes encompass many forms of media synthesis, including voice-swapping, text-to-speech, face-swapping, face-morphing, full-body puppetry, and lip syncing [49]. Moreover, recent progress in generative AI has enabled the cre-

ation of deepfakes using only text prompts, rather than requiring a data set of training images depicting the target individual [53, 72, 93]. While deepfake technology has potentially benevolent applications in creativity, accessibility, and entertainment [13, 19, 30, 31, 89], it has also been used to spread disinformation, commit fraud (e.g., phishing), and non-consensually generate intimate imagery [2, 15, 20].[1] The latter has commonly been termed "deepfake pornography," but following evolving terminology around image-based sexual abuse [58], we refer to it in this paper AI-generated non-consensual intimate imagery (AIG-NCII).[2]

Current technical research around deepfakes has predominantly focused on developing generative AI systems capable of synthesizing such content, including face-swapping [64, 97] and text-to-video systems [43, 78, 96], detection methods [12, 22, 98], as well as strategies to disrupt their generation [76]. However, research on attitudes of the general public towards deepfakes is far more nascent. A large body of literature and theory in information systems and HCI has underscored the importance technology acceptance — by individuals and by society — on technology use (and misuse) [46, 54]. Thus, this research seeks to bridge the gap between the technically possible (e.g., the academic research cited above) and the public acceptance of different uses of the technology. As computer security and privacy researchers, we are particularly interested in adversarial contexts, e.g., the generation of AIG-NCII. Hence, we ask: **What are people's attitudes toward the hypothetical non-consensual creation, sharing, and/or seeking out of deepfakes depicting them?** Decomposing this question, we ask specifically:

RQ1: How do attitudes differ depending on what is depicted: AIG-NCII vs. non-consensually created content depicting *non-sexual* acts?

---

[1] Intimate imagery refers to "images and videos of people who are naked, showing their genitals, engaging in sexual activity or poses, or wearing underwear in compromising positions" [80].

[2] AIG-NCII is our preferred term because it emphasizes the non-consensual nature of the images and is more widely applicable to the range of technologies that can be used to create such images.

RQ2: How do these attitudes differ depending on contextual factors: who is creating the media and for what purpose?

RQ3: How do attitudes related to sexual (a) consent and (b) content influence these attitudes?

RQ4: How does gender influence these attitudes?

To answer these questions, we conducted a vignette-based survey of 315 individuals to assess attitudes towards different situations involving non-consensual synthetic media. This research elucidates contextual and individual factors that shape public acceptance of generative AI technology being used to construct deepfakes in addition to broader trends in attitudes and rationales. Through this work, we aim to inform future discourse regarding deepfakes, specifically AIG-NCII, in public, technical, legal, and policy spheres.

2   Background & Related Work

In 2017, a user named "deepfakes" posted synthetic videos of celebrities in sexual acts to Reddit [31, 49, 51]. Over 90,000 users subsequently joined an r/deepfake subreddit for creating and sharing similar content, drawing significant public attention before being banned by Reddit as "involuntary pornography" [73]. Online communities catalyzed the popular use of the term "deepfake" [31, 49, 51], and despite bans on mainstream social media platforms, AIG-NCII continues to be produced and circulated on dedicated forums [2, 84].

Image-Based Sexual Abuse (IBSA). AIG-NCII is one form of IBSA: the non-consensual creation, distribution, or threats made with intimate images [56, 57, 77]. Victim-survivors of IBSA often experience severe health consequences, such as post-traumatic stress disorder, anxiety, depression, and greater somatic burdens [7, 25, 44, 77]. IBSA harms are also social, e.g., isolation, lowered self-esteem, trust issues, and unhealthy coping mechanisms [7, 56]. Victim-blaming attitudes are prevalent when seeking support or justice after IBSA [33], and obstruct help-seeking [59, 66]. IBSA falls under a broader umbrella of technology-facilitated gender-based violence [23]. As with other gender-based violence, victim-survivors of IBSA are predominantly, though not exclusively, women [2, 24].

IBSA and AIG-NCII are growing global issues [34]. Policy on IBSA is sparse in most countries [1, 92]; in the US specifically, legal scholars have called for legislation to sufficiently address its harms [17, 21, 37]. Understanding public attitudes about synthetic media, specifically AIG-NCII, can inform better policies on this emergent form of IBSA.

Public attitudes about AIG-NCII. Early research found significant public concern about non-sexual deepfake creation and dissemination [39], but less if created for entertainment, humor, or with consent and traceability [52, 63].

Regarding AIG-NCII, i.e., sexual deepfakes, prior work has primarily focused on attitudes around criminality and perceived harm to victim-survivors [32, 51, 86]. Kugler and Pace found that individuals in the UK perceived significant harms from and strongly favored criminalization of sexual and non-sexual deepfakes [51]. Further, videos being labeled as fake did reduce the perceived harm of non-sexual deepfakes, but did not for AIG-NCII [51]. Fido et al. study AIG-NCII while varying the identity of the target, finding that deepfakes of celebrities were perceived as less criminal and less harmful, especially for celebrities who are men [32]. This work also found that creation of deepfakes for personal sexual gratification was viewed as less harmful and criminal than sharing. Finally, in Umbach et al.'s study across ten countries, awareness of AIG-NCII was low overall, but surveyed individuals believed victims had a right to be upset [86]. Men in this study also reported more perpetration and victimization.

We combine elements from prior work on non-sexual deepfakes and AIG-NCII to systematically study *acceptance* (vs. criminality or harm) of the use of generative AI technology to create different types of deepfakes. Specifically, we extend [51] to compare AIG-NCII with not-exclusively-harmful deepfake actions (RQ1): saying something – which is ambiguous regarding sexuality or harmfulness – and playing a sport – ostensibly, a neutral action. We make these comparisons across five disambiguated actions involving deepfakes: creating, private sharing, public sharing, resharing, and seeking out. Additionally, we explore the role of contextual factors (RQ2) such as *intent* of the creator; a factor not explored in prior work on AIG-NCII despite the fact that intent is a factor in existing laws that can be applied to deepfakes and image-based sexual abuse [18] and the fact that prior work on non-sexual deepfakes finds that intent affects the general public's attitudes toward acceptability [52, 63]. As a second contextual factor, we further explore the relationship between the creator and subject; we explore the role of intimate partnership while prior work explored, and found relevant, celebrity status [32]. We further explore the impact of individual factors on these attitudes. We select individual factors found relevant in prior work on offline sexual abuse such as sexual consent attitudes [45] but which have been unexplored in the context of deepfakes and AIG-NCII (RQ3); as well as individual factors found relevant in prior work on AIG-NCII criminality perceptions such as gender [86] (RQ4).

Finally, as noted by Fido et al. [32], prior work lacks qualitative exploration of *why* respondents held particular opinions. In our work, we collect and analyze qualitative data on attitudes toward the acceptability of creating AIG-NCII and other synthetic media.

Deepfake community attitudes. Research has examined pro-deepfake views among Reddit users [36] and on MrDeep-Fakes [84], as well as positive attitudes but misuse concerns in a deepfake tool's open-source community [91].

3 Methodology

We conducted a survey of 315 U.S. Prolific respondents (survey instrument provided in the extended arXiv version of this paper [11]). Our Institutional Review Board (IRB) found our study to be exempt and we followed the ethical considerations as described in Section 3.4.

3.1 Survey structure

Consent. The survey began with a description of generative AI and its capacity to generate realistic-looking but fake media. We chose to avoid using "deepfake" given potential priming effects (e.g., about political disinformation). Respondents then were told survey structure and asked to consent.

Vignettes. We used vignettes—short descriptions of hypothetical scenarios—to solicit respondents' attitudes about AIG-NCII. Vignettes are common in security and privacy studies to elicit reactions [28, 55, 62] and can approximate real-world behaviors [41]. Drawing on the theory of contextual integrity [65], each vignette described generative AI being used to create a video of the respondent without their knowledge, varying three factors:

(1) action varies sexual explicitness, from unambiguously sexual behavior ('performing a sexual act') to non-sexual ('playing a sport') to ambiguous ('saying something'). This factor corresponds to RQ1.

(2) creator varies the relationship between the media maker and participant, either 'an intimate partner' or 'a stranger.' This corresponds to RQ2 and complements prior work [32, 51] exploring other relationships (e.g., of a celebrity).

(3) intent varies the creator's motivation, representing motivations reported by prior work [31, 89]: 'harming you,' 'entertainment,' and 'sexual pleasure,' also corresponding to RQ2.

One such vignette reads: "Imagine that an intimate partner uses generative AI to create a synthetic video of you playing a sport for the purpose of entertainment. Assume that you are unaware of the video's creation and existence." We employed a 2 (creator) × 3 (action) × 3 (intent) full-factorial design to construct 18 vignettes (see Table 1). The six vignettes where action was 'performing a sexual act' constitute cases of AIG-NCII. Other vignettes, such as V8, are not necessarily AIG-NCII but may still be sensitive. Each respondent was randomly assigned three vignettes to mitigate survey fatigue [68]. For each vignette, respondents rated the acceptability on a 5-point Likert scale from "Totally unacceptable" to "Totally acceptable"; for ratings other than "Neutral", they also wrote a short open-ended rationale about their choice.

Prior work found initial evidence [32, 51, 86] or hypothesized [81, 100] that acceptability may vary across behaviors. Thus, we assess acceptability for five AIG-NCII behaviors:

(1) creation of the video

ID	creator	action	intent
V1	an *intimate partner*	performing a *sexual act*	*entertainment*
V2	an *intimate partner*	performing a *sexual act*	*harm*ing you
V3	an *intimate partner*	performing a *sexual act*	*sexual pleasure*
V4	an *intimate partner*	playing a *sport*	entertainment
V5	an *intimate partner*	playing a *sport*	*harm*ing you
V6	an *intimate partner*	playing a *sport*	sexual pleasure
V7	an *intimate partner*	*saying something*	entertainment
V8	an *intimate partner*	*saying something*	*harm*ing you
V9	an *intimate partner*	*saying something*	sexual pleasure
V10	a *stranger*	performing a *sexual act*	*entertainment*
V11	a *stranger*	performing a *sexual act*	*harm*ing you
V12	a *stranger*	performing a *sexual act*	*sexual pleasure*
V13	a *stranger*	playing a *sport*	entertainment
V14	a *stranger*	playing a *sport*	*harm*ing you
V15	a *stranger*	playing a *sport*	sexual pleasure
V16	a *stranger*	*saying something*	entertainment
V17	a *stranger*	*saying something*	*harm*ing you
V18	a *stranger*	*saying something*	sexual pleasure

Table 1: The ID and contextual details of creator, action, and intent of each vignette. The italicized portions of the contextual details are the shorthand descriptions of the vignettes used in the paper text, e.g., V1 - intimate partner/sexual act/entertainment. The highlighted vignettes are AIG-NCII.

(2) private_sharing by the creator, e.g., in a group chat

(3) public_sharing by the creator, e.g., posting it on Reddit

(4) resharing, publicly, by someone who received the video from the creator

(5) seeking_out by someone with whom it was not shared, e.g., searching online by a description of the video

Sexual Consent Scale-Revised. To answer RQ3a about the role of attitudes towards sexual consent, we use two validated subscales from the Sexual Consent Scale-Revised (SCS-R) [45] (included in the extended arXiv version [11]) : SCS-R2 measures attitudes toward establishing consent, and SCS-R4 measures agreement with sexual consent norms based on relationship status and sexual activity. These subscales were selected over others from the SCS-R as our focus was on respondents' attitudes rather than self-reported behaviors.

Genuine Intimate Imagery (GII) and NDII Attitudes. To answer RQ3b about attitudes towards sexual content, we assessed attitudes on intimate media creation in intimate relationships. Paralleling the vignettes, we also asked about four scenarios involving non-consensual distribution of intimate images (NDII): (1) private sharing and (2) public sharing by the intended recipient, as well as (3) public sharing and (4) seeking out by someone who was *not* the intended recipient.

Demographics. The survey concluded with demographic questions, including gender (RQ4).

3.2 Respondents

We used power analysis to determine the required number of respondents for constructing our regression models with the ability to observe small-to-medium effects. We recruited 335 Prolific respondents who were over 18, lived in the US, and had over 95% approval on Prolific. 20 respondents who did not pass a Pew attention check question [14] or provided incoherent open-ended responses were excluded. The survey took an average of 15 minutes to complete. We compensated respondents $3, which we calculated based on our average pilot test length (12 minutes) and a rate of $15/ hour. 156 respondents were women, 150 were men, 6 were non-binary, 2 were agender, and 1 preferred not to say. Further demographic information is presented in Appendix A.

3.3 Data analysis

Quantitative analysis. Given that the dependent variable was a categorical Likert scale measuring acceptability judgments, and we aimed to include both fixed and random effects as independent variables, we analyzed respondents' attitudes using cumulative link mixed models (CLMMs). We built five CLMMs, one for each of the dependent variables concerning the synthetic video described in the vignettes, listed above. Each model included the same six independent variables. The first three were the vignette factors (creator, action, intent) (RQ1 & RQ2). For RQ3a, we included participant scores on the two SCS-R subscales. To evaluate potential co-linearity between variables, we tested the correlation between scores on the SCS-R subscales. Finding only a weak Spearman's correlation coefficient of -0.3 [3], we proceeded with including both subscales as distinct dependent variables.

Additionally, each model included one context-relevant independent variable capturing attitudes towards similar situations involving GII and NDII (RQ3b). For example, the model for creation included attitudes towards the creation of GII within an intimate partnership as an independent variable and the model for private_sharing included attitudes towards the indented recipient of GII sharing it privately outside the relationship, without consent. During initial analysis, we decided to bucket these attitude items into "unacceptable" and "not unacceptable" to increase our statistical power. Lastly participant gender (bucketed into men and minoritized genders, see below) was included to address RQ4.

AIG-NCII is a form of image-based sexual abuse and tech-facilitated gender-based violence, which is predominantly, though not exclusively, perpetrated by cisgender men targeting cisgender women, transgender people, and/or non-binary people [23, 24, 56, 57, 94]. While research continues to investigate gendered proportions of perpetration and victimization—one report finds that most online AIG-NCII targeted women [2], another report finds that men were more likely to report AIG-NCII victimization than women [86]—

attitudes are nevertheless informed by the broader dynamics of gender-based violence. Thus, mens' attitudes of AIG-NCII may differ from the attitudes of people who are not men. In order to increase statistical power, we grouped people who were not men together, i.e., women, agender, or non-binary individuals and refer to this group as "marginalized genders."[3] Further, we only had 8 respondents who self-identified as agender, or non-binary; we bucketed them with women to include their responses in our quantitative analyses, rather than dropping the responses entirely. Additionally, we ran statistical models for 'women' and 'men', excluding participants outside this gender binary, which are similar and lead to the same conclusions (see Appendix D).

To further examine the contextual factors' effect on acceptability (RQ2), another CLMM was built by adding interactions terms between intent and action as well as intent and creator to the original model for creation. To examine the effect of participant gender on attitudes towards synthetic imagery (RQ4), five additional models were built by expanding the original models to include interactions terms between gender and each vignette factor. Of the expanded models, only the creation model showed statistically significant interaction effects ($p < 0.05$) and thus was selected for further analysis. To compare acceptability across the actions of creation, private_sharing, public_sharing, resharing, and seeking_out, another model was built with acceptability rating as the dependent variable and these actions as the independent variable.

Qualitative analysis. We analyzed respondents' open-text rationale for their acceptability rating for the creation of the synthetic video using a coding reliability approach [9]. The dataset was divided into two subsets, justifications for and against acceptability. Two researchers familiarized themselves with all rationales and generated an initial set of codes. The researchers compared and discussed codes to establish a final codebook (Appendix C). In line with qualitative research perspectives on the limitations of multiple coders [4, 60], a single researcher performed the entire coding process for consistency and to preserve interpretive nuance [26]. A second researcher reviewed the codebook as well as 50 random responses from each subset in order to balance researcher subjectivity with thoroughness [90].

3.4 Other considerations

Ethical considerations. This study was deemed exempt by our IRB. However, ethical considerations extend beyond regulatory compliance [8]. As vignettes describe non-consensual creation and sharing of intimate imagery, we were concerned about potential harm from placing respondents into hypothetical victimization scenarios, especially for those who have experienced image-based sexual abuse or sexual violence.

---

[3]In our survey, we did not ask whether respondents were transgender, so our sample of men includes transgender and cisgender men.

Consulting subject-area experts with training in clinical psychology and sexual trauma, we took the following steps for harm reduction: (1) surfacing in the consent form that the vignettes described synthetic media being created of the respondent, (2) asking for re-consent after defining generative AI, (3) including 'prefer not to answer' option for all questions about intimate images, and (4) including contact information for IBSA support organizations at the end of the survey. We also provided support resources for members of the research team who analyzed open-ended survey responses.

Positionality statement. Recognizing the inherent subjectivity in research, we acknowledge that our positionality as researchers shapes our approach to this work [6, 10, 42]. We bring varied perspectives informed by our distinct social, cultural, disciplinary, and ideological contexts. Our research team consists of three cisgender women and one cisgender man who are all researchers in security and privacy. As our team composition does not fully reflect the diversity of identities among our study respondents, there may be limitations in our thematic analysis and interpretation of the collected data.

Limitations. While surveys offer valuable insights, there are inherent limitations to using them. We prioritized reducing survey fatigue by pre-testing and piloting our survey. To minimize social desirability bias, we emphasized that each response about acceptability was based solely on the respondent's personal opinions. Our data is limited to the attitudes and justifications respondents were willing to report.

Crowdworking platforms offer access to large and diverse populations and are frequently used to elicit security and privacy attitudes [28, 75, 88]; we chose Prolific for its higher data quality compared to other platforms [67, 70]. Anticipating that attitudes towards AIG-NCII vary by country, we chose to recruit solely in the US, which likely limits generalizability.

As noted in Section 3.3, our survey instrument did not record transgender identities. As a result, our analysis may not fully capture the experiences of transgender individuals.

Additionally, as a formative study, we chose to explore specific factors (e.g., gender, contexts) rather than formulate uninformed hypotheses.

4 Results

To quantitatively analyze the 315 survey responses, we built eight CLMMs (see Section 3.3). The complete regression results for five, including the odds ratio (OR), confidence interval, and *p*-value range for each independent variable, are in Table 2 (see the extended arXiv version for visualization [11]). Where models with interactions are used (Table 3 and Table 4), only the models for creation had significant interaction terms and thus were selected for analysis.

Additionally, we conducted thematic analysis of the 861 open-response explanations of why participants found the creation of synthetic media in each vignette either acceptable or

unacceptable. Aligned with qualitative methods, our analysis aimed to surface general themes about participants' attitudes, rather than quantify their prevalence. Accordingly, we report the appearance of themes using the following terminology: a few (less than 25%), some (25-45%), about half (45-55%), most (55-75%), and almost all (75-100%). When providing participant quotes, we refer to each participant with the letter 'P' followed by their unique participant number and specify the vignette they were responding to. Visualizations for the distributions of codes over vignettes and actions are available in the extended arXiv version of this paper [11]. In some figures and this section, vignettes are referenced by their ID (e.g., V5) and the factor description creator/action/intent (see Table 1).

In our results, we use *synthetic* media to refer to media that is AI-generated, e.g., deepfakes, and *AIG-NCII* to refer to synthetic media that are specifically intimate imagery.

4.1 General Attitudes (RQ1)

People generally found the creation of synthetic media unacceptable, with a median percentage of somewhat or totally unacceptable ratings across all scenarios of 89.54%. They perceived any sharing of these media as even more unacceptable: 94.39% for private_sharing, 94.44% for public_sharing, 94.22% for resharing. Attitudes were more mixed regarding seeking_out such media, however (52.78%). The results of the regression examining the acceptability rating as the dependent variable with these actions as the independent variable, support these results statistically (see Table 6 in Appendix B for full results): Across scenarios and controlling for within-subject variation we observe that private_sharing (OR = 0.47, $p < 0.001$), public_sharing (OR = 0.26, $p < 0.001$), and resharing (OR = 0.42, $p < 0.001$) are significantly less acceptable than creation (the reference level). seeking_out (OR = 5.43, $p < 0.001$) is significantly more acceptable than creation.

Figure 1 illustrates these results visually, depicting perceived acceptability across creation, private_sharing, public_sharing, resharing, and seeking_out for all vignettes. The rightmost column (seeking_out) exhibits far more variance in attitudes than the columns to the left, although these variances differ depending on the depicted action, as we investigate next.

AIG-NCII perceived as less acceptable than other synthetic media not depicting sexual acts. While people broadly found creation and any form of sharing of synthetic media unacceptable, this was particularly true for AIG-NCII (RQ2). Across creation, private_sharing, public_sharing, and resharing contexts, scenarios in which the action was playing a sport or saying something, as opposed to performing a sexual act, were rated as more acceptable by participants (OR > 7, $p < 0.001$ for all models in Table 2).

Turning again to Figure 1, we observe this effect clearly.
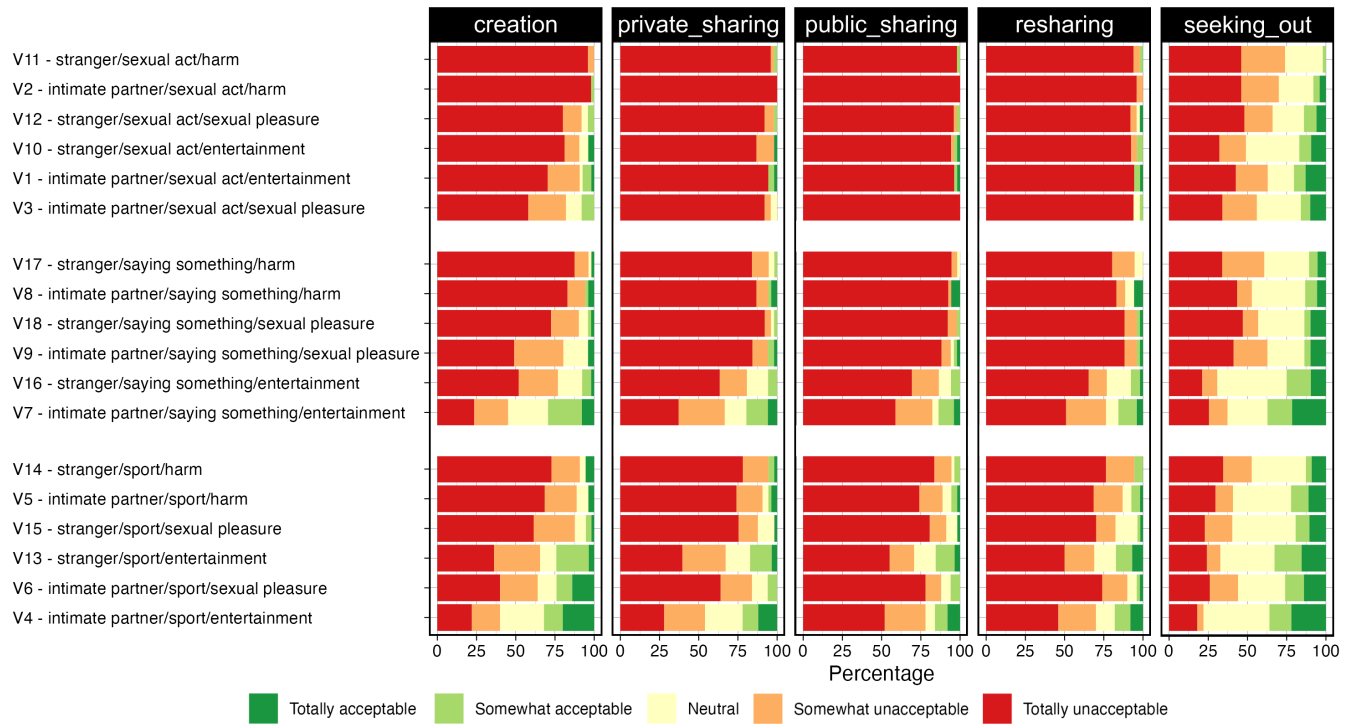
Figure 1: Respondents' perceptions of acceptability across all vignettes; each vignette is defined by the creator / action / intent. Vignettes are grouped by action and ordered (from bottom to top) by increasing unacceptability of creation.

Regarding creation, the least unacceptable scenario depicting a sexual act was V3 – an intimate partner non-consensually creating synthetic media of the participant engaged in a sexual act for their sexual pleasure – 82% of respondents found this scenario to be somewhat or totally unacceptable. [4] The most accepted scenario depicting the participant saying something (V7) – an intimate partner non-consensually creating synthetic media of the participant saying something for entertainment – was considered unacceptable by about half of participants (45.1%). The most acceptable scenario in our entire survey (V4), which depicted an intimate partner non-consensually creating synthetic media of the participant playing a sport was considered unacceptable by just a third (32%) of participants.

seeking_out AIG-NCII was also viewed as less acceptable than seeking_out other forms of synthetic content (OR > 3, $p < 0.001$; Table 2). However, when comparing seeking_out AIG-NCII to creating it, it is still more acceptable than creation as illustrated by Figure 2.

Portrayed action relates to perceived harm. When explaining their perception of a scenario, some participants remarked on potential harm to their reputation or lack thereof to explain why they viewed creation as acceptable or unacceptable. Lack of harm was the most common reason for finding synthetic media creation acceptable, typically when that media depicted

---

[4] The potential for flattery within a relationship (see Section 4.2) may explain why this was lower than the median across all vignettes.
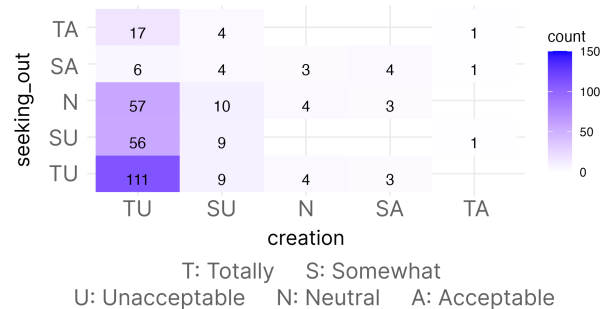
Figure 2: Heatmap of acceptability for creation and seeking_out when the action is performing a sexual act.

the subject playing a sport. For example:

There is nothing sexual. . . that i woul[dn']t want the public to know/see (P50, V13 - stranger/sport/entertainment).

On the other hand, when discussing AIG-NCII or depictions of them saying something they did not, some participants remarked on the potential harms of that content:

Sexual act will tarnish my image in the society (P193, V10 - stranger/sexual act/entertainment).

AI can seem realistic. Whatever they have me saying could be used against me in a variety of situations (P32, V16 - stranger/saying something/entertainment).

	creation	private_sharing	public_sharing	resharing	seeking_out
Intercepts					
Totally unacceptable | Somewhat unacceptable	5.13 [0.42, 62.94]	1.19 [0.29, 115.33]	29.21 [0.86, 987.2]	1.02 [0.04, 24.7]	0.03* [0, 0.8]
Somewhat unacceptable | Neutral	29.38** [2.35, 367.77]	5.82 [0.29, 115.33]	122.87** [3.5, 4318.76]	4.72 [0.19, 114.86]	0.16 [0.01, 4.42]
Neutral | Somewhat acceptable	99.64*** [7.82, 1269.53]	18.57 [0.93, 370.82]	289.97** [8.06, 10433.52]	15.28 [0.62, 374.82]	5.29 [0.19, 145.78]
Somewhat acceptable | Totally acceptable	375.02*** [28.62, 4913.68]	83.63** [4.08, 1713.2]	1481.83*** [38.75, 56663.78]	70.09* [2.78, 1767.14]	25.60 [0.92, 710.87]
Controlled IVs					
creator (Intimate partner)	3.24*** [2.23, 4.71]	1.69* [1.13, 2.55]	1.47 [0.9, 2.4]	1.00 [0.65, 1.53]	1.11 [0.8, 1.53]
action (Sport)	13.39*** [7.96, 22.52]	34.72*** [16.76, 71.92]	66.61*** [22.75, 19504]	32.36*** [15.12, 69.25]	7.26*** [4.73, 11.15]
action (Saying something)	5.44*** [3.27, 9.05]	11.01*** [5.45, 22.23]	19.49*** [6.91, 54.94]	12.47*** [5.92, 26.29]	3.40*** [2.21, 5.22]
intent (Entertainment)	18.92*** [11.03, 32.46]	11.49*** [6.59, 20.05]	10.57*** [5.39, 20.73]	5.51*** [3.18, 9.56]	4.94*** [3.25, 7.49]
intent (Sexual pleasure)	7.42*** [4.42, 12.47]	1.35 [0.77, 2.37]	1.15 [0.58, 2.28]	0.92 [0.52, 1.63]	1.37 [0.92, 2.04]
Uncontrolled IVs					
Gender (Man)	2.45*** [1.45, 4.15]	2.12** [1.21, 3.7]	1.77 [0.88, 3.55]	1.41 [0.75, 2.66]	1.51 [0.76, 2.99]
GII & NDII attitudes (Unacceptable)	0.21* [0.05, 0.84]	0.08* [0.01, 0.4]	0.09** [0.02, 0.41]	0.01*** [0, 0.05]	0.01*** [0.01, 0.03]
SCS-R2	0.53*** [0.39, 0.72]	0.55*** [0.4, 0.77]	0.64* [0.42, 0.96]	0.76 [0.52, 1.1]	0.73 [0.48, 1.1]
SCS-R4	1.06 [0.82, 1.36]	1.10 [0.84, 1.44]	1.30 [0.92, 1.82]	1.27 [0.93, 1.72]	1.14 [0.82, 1.59]

Table 2: Results from regressions exploring the relationship between scenario acceptability (first row, intercepts), contextual factors (second row, controlled IVs), and personal factors (third row, uncontrolled IVs). Each column represents the output of one regression model. Numeric cells list the odds ratio (OR) and the 95% confidence interval. Reference levels: creator (stranger), action (sexual act), intent (harm), gender (marginalized genders), GII & NDII attitudes (acceptable). Significance of OR: $p < 0.05 = $ *, $p < 0.01 = $ **, and $p < 0.001 = $ ***.

Further, when the action was performing a sexual act, a few participants also observed that the creation of AIG-NCII wrong because — even if synthetic — the images violated the sanctity of their bodies, e.g.:

It's a violation of my body and it is disrespectful (P49, V10 - stranger/sexual act/entertainment).

I feel it's unacceptable to manipulate my image in such a way - my body and how it looks belongs to me (P195, V1 - intimate partner/sexual act/entertainment).

Finally, while we only asked respondents to explain their judgements of (un)acceptability relating to media creation (Section 3.1), some mentioned the stage of media production (e.g., creation vs. any form of sharing) influenced the likelihood of harm and thus their perception of acceptability:

It's not harming me or blackmailing me or anything. As long as it doesn't get shared I think it's ok (P163, V3 - intimate partner/sexual act/sexual pleasure).

Some respondents call on morality, legality, and privacy to explain the unacceptability of synthetic media. A few participants justified the creation of synthetic media depicting them as unacceptable because it was amoral or unethical to create fake content without the subject's consent, e.g.,

This is a false representation of me and highly unethical (P204, V16 - stranger/saying something/entertainment).

I don't think it is right to use a person[']s identity to say things that they didn't say (P302, V16 - stranger/saying something/entertainment).

While not specifically speaking to amorality, a few expressed sentiments of disgust often associated in psychological literature with intuitive responses to moral violations [38]: that the creation of the content was 'gross' (P50), 'creepy' (P24), 'weird' (P74), or 'nasty' (P112). Such feelings were especially prevalent when the content was created by a stranger or the action depicted was incongruous with the intent (e.g., a stranger creating a video of someone playing a sport for sexual pleasure). We explore these variations based on contextual factors further in Section 4.2.

In a few other cases, participants referred to the creation of the media as illegal or compared it to a crime, despite

the fact that no federal legal protections currently exist on AIG-NCII [92]. Across all actions, a few participants called the act of creation slanderous, like P268 in response to V14 (stranger/sport/harm):

They are using faked info to harm me. This is slander.

When the action was saying something, the creation was often compared to libel or fraud, e.g.,

It seems like the equivalent of slander and fraud. If this were done in election ads, it would be disallowed/illegal (P253, V17 - stranger/saying something/harm).

[I]t is never acceptable to lie. I would sue for libel (P259, V7 - intimate partner/saying something/entertainment).

Specific to AIG-NCII, participants mentioned crimes of sexual violence,

This scenario is harmful and akin to some form of sexual ha[r]assment or assault, especially done without knowledge (P212, V2 - intimate partner/sexual act/harm).

Finally, a few respondents called the creation of synthetic media of them a privacy violation, e.g.:

This completely violates my sense of privacy (P10, V2 - intimate partner/sexual act/harm).

Creating an image of a person without their knowledge is a violation of privacy (P170, V6 - intimate partner/sport/sexual pleasure).

This attitude appeared relatively evenly and similarly in rationales across all actions.

4.2 Role of contextual factors (RQ2)

Consistent with the theory of contextual integrity [65], we found that contextual factors strongly influenced both respondents' ratings of acceptability and their rationales.

It is more acceptable for intimate partners to create synthetic media than strangers, but only if they do not intend harm. We observe from Table 2 that across all scenarios, when the content creator was an intimate partner as opposed to a stranger, participants were more likely to find the creation (OR = 3.24, $p < 0.001$; Table 2) as well as the private_sharing (OR = 1.69, $p = 0.01$; Table 2) of the synthetic imagery more acceptable (RQ2). However, when we consider interactions with the intent of the synthetic media (Table 3), we observe that there is no longer a significant relationship between creator and acceptability of creation and that there are three significant interactions between: (1) creator being an intimate partner and intent being entertainment (OR = 2.83, $p = 0.036$; Table 3), (2) creator being an intimate partner and intent being sexual pleasure (OR = 3.76, $p = 0.009$; Table 3), as well as between (3) action being playing a sport and intent being sexual pleasure (OR = 0.08, $p = 0.002$; Table 3), which

		OR; Confidence Interval
Intercepts	Totally unacceptable | Somewhat unacceptable	7.51; [0.42, 134.86]
	Somewhat unacceptable | Neutral	47.58; [2.61, 867.1]**
	Neutral | Somewhat acceptable	171.35; [9.26, 3169.4]***
	Somewhat acceptable | Totally acceptable	665.83; [35.15, 12613]***
Controlled IVs	creator (Intimate partner)	1.38; [0.62, 3.04]
	action (Sport)	48.94; [11.43, 209.59]***
	action (Saying something)	9.9; [2.24, 43.77]**
	intent (Entertainment)	13.14; [2.86, 60.3]***
	intent (Sexual pleasure)	20.82; [4.53, 95.72]***
Uncontrolled IVs	Gender (man)	2.64; [1.53, 4.57]***
	GII & NDII attitudes (Unacceptable)	0.19; [0.04, 0.8]*
	SCS-R2	0.51; [0.37, 0.71]***
	SCS-R4	1.08; [0.83, 1.4]
Interaction Terms	creator (Intimate partner) & intent (Entertainment)	2.83; [1.07, 7.5]*
	creator (Intimate partner) & intent (Sexual pleasure)	3.76; [1.38, 10.2]**
	action (Sport) & intent (Entertainment)	0.72; [0.15, 3.56]
	action (Saying something) & intent (Entertainment)	1.68; [0.33, 8.66]
	action (Sport) & intent (Sexual pleasure)	0.08; [0.02, 0.4]**
	action (Saying something) & intent (Sexual pleasure)	0.19; [0.04, 1.01]

Table 3: Results from a single regression exploring the relationship between the acceptability of creation (first row, intercepts), contextual factors (second row, controlled IVs), personal factors (third row, uncontrolled IVs), and interactions between intent and creator or action (fourth row, interaction terms). Reference levels: creator (stranger), action (sexual act), intent (harm), gender (marginalized genders), GII & NDII attitudes (acceptable). Significance of OR: $p < 0.05$ = *, $p < 0.01$ = **, and $p < 0.001$ = ***.

we address later in this section. Thus, our interaction model demonstrates a more nuanced answer to RQ2. The main effect we observed in our original modeling for creation (without interactions) – that intimate partners creating synthetic media is more acceptable – was driven by attitudes that intimate partners creating synthetic media for non-harmful purposes is more acceptable. That is, if the creator is an intimate partner and the intent is entertainment (OR = 2.83, $p = 0.036$; Table 3) or sexual pleasure (OR = 3.76, $p = 0.009$; Table 3) the media creation is more acceptable. However, intimate partners creating media for the intent to harm is no more acceptable than a stranger doing so.

Intimate partner trust related to explanations of (un)acceptability. Some explanations for acceptability, like P211's response to V1 (intimate partner/sexual act/entertainment), reflected trust in a partner enabling ac-

ceptable creation:

I feel if we are intimate, we're already engaging in similar acts. It's all in good sexual fun, as long as they don't distribute it or show anyone else.

This exhibits a belief that an intimate relationship permits intimate media creation within it, whereas no such trust exists in relationships with strangers, increasing feelings of violation:

The idea of somebody I don't know generating porn of me is insanely creepy (P24, V12 - stranger/sexual act/sexual pleasure)

On the other hand, some explanations for unacceptability stated that the creation *violated* intimate partner trust rather than being acceptable because of it, e.g.,

I think this is just as worse because there is supposed to be a trust between people who are intimate and they completely broke that trust (P142, V3 - intimate partner/sexual act/sexual pleasure).

About half of the rationales exhibiting this attitude were in response to the creation of synthetic media of sexual acts.

A few were flattered by the creation of material for sexual fantasy within an intimate partnership. In scenarios where synthetically generated media was created for sexual gratification by an intimate partner, a few participants reported feelings of being flattered by its production, e.g.,

The content she generated sounds cool and indicates she's attracted to me (P65, V6 - intimate partner/sport/sexual pleasure).

I don't care what my intimate partners choose to do. I would be flattered (P65, V9 - intimate partner/saying something/sexual pleasure).

A few noted that they couldn't control the sexual fantasies of others, regardless of whether they were in a relationship:

I don't particularly like that and I would prefer they don't do it, but I can't stop them from fantasizing about me in their own head. I can't stop them from writing down their fantasies on paper or drawing a picture (P188, V12 - stranger/sexual act/sexual pleasure).

While others expressed that, in the context of an intimate relationship, they would prefer to engage in their partner's fantasy instead:

It's a bit bizarre and strange. I'd rather I actually perform this act instead of a fake AI version of me doing so (P165, V1 - intimate partner/sexual act/entertainment).

Intent impacts acceptability ratings differently depending on stage in the media pipeline. We observe from Table 2 that regardless of the creator of the media, respondents rated as

more acceptable those scenarios where synthetic videos were created, shared, and sought out for entertainment vs. with intent to harm ($OR > 4$, $p < 0.001$; Table 2). Respondents also found creation of synthetic videos with the intent of bringing the creator sexual pleasure more acceptable than creation with the intent to harm the subject. However, respondents did not rate the acceptability of any form of sharing or seeking_out synthetic videos created with the intent of sexual pleasure differently from the acceptability of sharing or seeking_out synthetic videos created with the intent to harm.

Incongruent actions and intentions increase unacceptability. Considering our interaction model, we find that these results hold but observe a further effect: incongruence between the action and the intent – even for actions and intents viewed as generally more acceptable – reduce attitudes of acceptability. For example, while creating media depicting the subject playing a sport was overall more acceptable than depicting them engaged in a sexual act and depictions of any action for sexual pleasure were more acceptable than depictions for harm, depicting someone playing a sport with the intent of sexual pleasure was less acceptable than depicting a more congruous action (saying something, a sexual act) with the same intent. A few participants shared explanations for the (un)acceptability of synthetic media creation that support this finding, for example:

That's really creepy! It just grosses me out, even if it's just sports. (P25, V15 - stranger/sport/sexual pleasure)

4.3 Role of sexual consent & content attitudes (RQ3)

Attitudes toward establishing sexual consent offline relate to attitudes toward AI media generation and sharing. We used the second subscale from the SCS-R to measure attitudes towards establishing sexual consent [45] and answer RQ3a. Those who scored higher on SCS-R2, indicating more positive attitudes toward establishing sexual consent, were less likely to rate non-consensual creation, private_sharing or public_sharing of synthetic content as acceptable ($OR < 0.7$, $p < 0.005$ for these models; in Table 2).

The most common explanation for finding synthetic media creation unacceptable is lack of consent. For example, P19 remarked in response to V3 (intimate partner/sexual act/sexual pleasure) that:

No content should be made in someone else's likeness without their consent.

The fourth SCS-R subscale measures attitudes towards consent norms specifically in the context of relationships and sexual activity [45]. Scores on this subscale did not significantly affect any models.
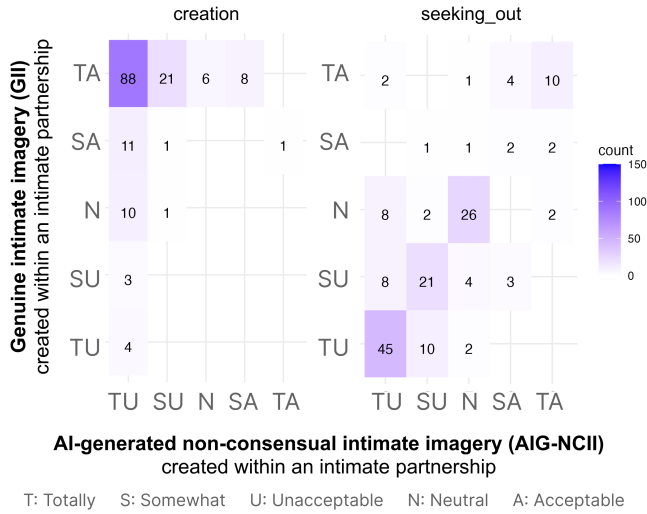
Figure 3: Heatmaps comparing acceptability of creation and seeking_out for AIG-NCII to similar actions for GII also created in an intimate relationship. See the extended arXiv version [11] for heatmaps including all forms of sharing.

Attitudes toward consensually-created genuine intimate imagery as well as NDII correlate with acceptance of synthetic videos including AIG-NCII. In addressing RQ3b, we sought to understand whether and how attitudes toward genuine, consensually-created intimate imagery related to attitudes toward synthetic, non-consensually created media.

Those who found consensual creation of genuine intimate imagery (GII) in an intimate relationship (somewhat or completely) unacceptable were also less likely to find non-consensual, synthetic creation of media depicting them acceptable, regardless of the act depicted (OR = 0.21, p = 0.028; Table 2). Those who found further sharing of GII without the original sender's consent – i.e., non-consensual distribution of intimate imagery or NDII – unacceptable were also less likely to find sharing of synthetic videos depicting them acceptable (OR < 0.1, p < 0.05 for private_sharing, public_sharing, and resharing; Table 2). Finally, those who considered seeking out NDII unacceptable were less likely to find seeking_out synthetic videos acceptable (OR = 0.01, p < 0.001; Table 2).

In Figure 3, we observe that over three fourths of participants who responded to a vignette involving AIG-NCII in the context of an intimate relations found consensual GII creation within an intimate partnership totally acceptable (116/154), while none viewed non-consensual synthetic intimate media creation within an intimate partnership as totally acceptable. A key difference is that the GII creation scenario implies awareness and consent, while the synthetic media vignettes explicitly do not. Considering non-consensual sharing, a majority of respondents viewed private_sharing (140/153[5]), public_sharing (145/151), and resharing (139/153) as totally un-

---
[5]Denominators vary because some participants preferred not to answer certain questions about synthetic and/or genuine intimate imagery.

		OR; Confidence Interval
Intercepts	Totally unacceptable | Somewhat unacceptable	4.77; [0.33, 69.72]
	Somewhat unacceptable | Neutral	29.29; [1.97, 435.77]*
	Neutral | Somewhat acceptable	104.04; [6.89, 1570.59]***
	Somewhat acceptable | Totally acceptable	410.88; [26.53, 6363.7]***
Controlled IVs	creator (Intimate partner)	1.75; [1.04, 2.96]*
	action (Sport)	17.58; [8.06, 38.33]***
	action (Saying something)	10.68; [4.82, 23.65]***
	intent (Entertainment)	20.05; [9.39, 42.85]***
	intent (Sexual pleasure)	4.90; [2.32, 10.33]***
Uncontrolled IVs	Gender (man)	1.54; [0.43, 5.61]
	GII & NDII attitudes (Unacceptable)	0.2; [0.05, 0.84]*
	SCS-R2	0.52; [0.38, 0.72]***
	SCS-R4	1.08; [0.83, 1.41]
Interaction Terms	action (Sport) & Gender (Man)	0.77; [0.29, 2.02]
	action (Saying something) & Gender (Man)	0.32; [0.12, 0.88]*
	intent (Entertainment) & Gender (Man)	1; [0.38, 2.61]
	intent (Sexual pleasure) & Gender (Man)	2.36; [0.87, 6.43]
	creator (Intimate partner) & Gender (Man)	3.59; [1.71, 7.5]***

Table 4: Results from a single regression exploring the relationship between scenario acceptability for creation (first row, intercepts), contextual factors (second row, controlled IVs), personal factors (third row, uncontrolled IVs), and interactions between gender and contextual factors (third row, interaction terms). Reference levels: creator (stranger), action (sexual act), intent (harm), gender (marginalized genders), GII & NDII attitudes (acceptable). Significance of OR: p < 0.05 = *, p < 0.01 = **, and p < 0.001 = ***.

acceptable for both media types. There was less consensus on seeking_out non-consensually publicized synthetic and non-synthetic imagery, with only some (45/154) finding it totally unacceptable for both.

4.4 Role of gender (RQ4)

For quantitative analysis, we binned respondents by gender into men and marginalized genders (see Section 3.3). Across scenarios, men were more likely to rate the creation (OR = 2.45, p < 0.001; Table 2) and private_sharing (OR = 2.12, p = 0.009; Table 2) more acceptable than people with a marginalized gender.

Men view synthetic media depicting them engaged in a sexual act more acceptable than others. To further examine the role of gender identity in shaping attitudes towards non-consensual synthetic imagery creation, we performed an additional regression that included interaction terms between

participant gender and each vignette factor (Table 4). We observe that the main effect of gender is no longer significant, instead finding two significant interactions with gender. The first shows that, while participants viewed creation of synthetic videos of them saying something as more acceptable than a sexual act, people of marginalized genders were more likely to do so than men (OR = 10.71 for men vs. OR = 3.42 for marginalized genders, $p = 0.027$).

Participants who are men are more accepting of intimate partners creating synthetic videos depicting them. Secondly, we observe that, holding all other factors constant, men were more likely to rate the creation of synthetic media by an intimate partner more acceptable (OR = 1.77 for men vs. OR = 6.27 for marginalized genders, $p < 0.001$). Additionally, most participants who described the creation of AIG-NCII in an intimate partnership as being acceptable because it was a compliment or part of their partner's fantasy (as discussed in Section 4.2) were men.

5 Discussion

Overall, we find that creating, sharing, or seeking AIG-NCII is considered far less acceptable than creating, sharing, or seeking other forms of non-consensually-created synthetic media (RQ1: Section 4.1). Respondents were more accepting of intimate partners creating synthetic media of them than strangers, including AIG-NCII, but only when their intent in doing so was not to cause harm (RQ2: Section 4.2). Lack of consent was the most common reason respondents provided for why non-consensual creation of synthetic media, including AIG-NCII, was unacceptable. Our statistical models support this finding: positive attitudes toward sexual consent were inversely correlated with acceptance of non-consensual creation, sharing, or seeking_out of synthetic media of any kind (RQ3: Section 4.3). The second most common reason respondents gave for why creation was unacceptable was potential for harm, either reputational damage or bodily violation; conversely, the lack of potential for such harm was the most common reason among those who found creation acceptable. Men in particular were more accepting of synthetic media creation (RQ4: Section 4.4), especially by intimate partners. We hypothesize based on prior literature on perceptions of sexual reputation in the context of defamation law [5, 71, 79] and participants' open-text responses that this is likely due to differences in perception regarding reputation damage and creation as a form of compliment as well as, from a critical perspective [74], that men may be more accepting of such images if they have more power in a relationship. Respondents also expressed attitudes of unacceptability due to moral violations [38], including feelings of disgust, and privacy violation.

We focus the remainder of our discussion on implications for addressing the most unacceptable use of AI generative capabilities we find in our study, AIG-NCII, although we note that the implications are relevant to other synthetic media.

Distributed responsibility and individual deterrence. We believe it is important to understand the gap between the unacceptability of creation and sharing and the relative acceptability of searching for, and subsequently viewing, of AIG-NCII. Based on our results, we hypothesize that one contributing factor to the continued ubiquity of AIG-NCII is the broad acceptance of or neutrality toward searching for such content. The finding that searching for and viewing AIG-NCII is perceived as so acceptable suggests the harms entailed in AIG-NCII are not fully appreciated by many people. Yet as studies of the experiences of image-based sexual abuse victim-survivors and even legal cases note, viewing is a primary mechanism of harm for NCII: "there [is] a fresh intrusion of privacy when each additional viewer sees the photograph" [48].

Past works, although not written in the context of AIG-NCII, can provide possible explanations for this gap, which we encourage future research to explore in depth. As media scholar Lilie Chouliaraki concludes in her analysis of the viewing of violent imagery in television and online, "technology closes the moral distance between spectators and sufferers and . . . yet, at the same time, it fictionalizes suffering and leads spectators to indifference" [16]. Media scholar Charles Ess [29], in his foundational work *Digital Media Ethics*, argues that such indifferent online behavior in new media networks is due to "distributed responsibility," which refers to the idea that ethical responsibility for an act is distributed across an interconnected, online networks of actors, rather than being attached solely to a single individual [29, 87]. Ess contrasts this collective responsibility with the traditional western understanding of ethical responsibility as matter of individual agency. For example, an individual might never steal an album from a physical record store but may illegally download of music from the Internet. In this and many cases, he argues, individuals consider themselves part of an anonymous, undetectable online collective without fear of punishment.

Thus, a key question for future work is how to combat indifference towards the harm of viewing AIG-NCII. Deterrence messaging, such as keyword-based warnings in search engines or advertisements that inform the viewer about the harms of consuming AIG-NCII, could be used to target individuals' sense of ethical immunity. Emphasizing personal accountability within the collective space could disrupt feelings of distributed responsibly related to AIG-NCII. Such messaging is currently effectively used to deter viewing of child sexual abuse material [69] but further research is necessary to find effective approaches to deter AIG-NCII consumption.

Harms vs. rights When analyzing our data, we observed different classes of arguments for (and against) the unacceptability of AIG-NCII. At the highest level, we saw arguments focused on harms and arguments focused on rights. For example, some argued that creating AIG-NCII was acceptable

as long as no harms manifested, e.g., "It's not harming me or blackmailing me ... [a]s long as it doesn't get shared I think it's ok" (Section 4.1): a harms-based analysis. On the other hand, some argued that creating AIG-NCII was unacceptable, even if never shared, because it was a "violation of my body" (Section 4.1): a rights-based evaluation.

While prior work on AIG-NCII has primarily focused on harm perceptions [32, 51], these two categories of arguments — harms-based and rights-based — align with the vast literature in philosophy and psychology on how different people may center different values in moral decision making, e.g., see [50] for a survey aimed at the security and privacy community. Using the terminology from philosophy, those who consider AIG-NCII unacceptable because it can lead to harms are centering a utilitarianistic (consequentialist) perspective on ethics; those who consider AIG-NCII unacceptable because it violates an individual's rights even if no harms manifest are centering a deontological perspective.

While our findings surfaced a breadth of rights that participants believe are impacted by the creation and possible sharing of AIG-NCII, we focus on two below: the right to consent, which is baked into the definition of AIG-NCII, and, given the SOUPS community, the right to privacy.

AIG-NCII as a consent violation. To our knowledge, ours is the first work to surface qualitative perspectives on consent for AIG-NCII. Our findings (Section 4.3) suggest connections between understandings and norms around consent in different contexts. Grounded in the observed relationships among respondents' acceptability ratings, attitudes towards sexual consent, and their free response explanations, we speculate on the potential implications of these context connections: First, shaping or enforcing norms around sexual consent, or consent in general, could influence norms and behaviors related to non-consensual synthetic media. Consent education, which involves setting and modeling behavioral norms like asking for consent before interacting with another person's body or space, is one approach to establishing and enforcing norms around consent for all ages in both sexual and non-sexual contexts [35, 83]. Second, centering consent as a priority in policies and technical developments around deepfakes is warranted. A growing body of work provides useful frameworks for operationalizing consent in sociotechnical systems [47, 82, 99].

AIG-NCII as a privacy violation. Like consent, privacy is a fundamental right. While our survey instrument did not mention privacy at any point, some participants stated that the creation of the synthetic media would violate their privacy.

The fact that contextual factors such as who created the content and for what purpose influence perceptions of AIG-NCII acceptability in our study aligns with existing technology privacy theory on contextual norms [95] and integrity [65], which find that experiences of privacy violation are dependent on contextual factors including what information is being

shared, which actors are involved, and the purpose of the information sharing. Thus, frameworks of privacy as contextual integrity may be one useful component of future policies about AIG-NCII.

At the same time, existing frameworks and technological conceptions of privacy often focus on *data* privacy. Yet, as technological capabilities continue to develop, technologists must increasingly contemplate how to measure and protect a more nebulous privacy right: to representational privacy. Creating AIG-NCII may involve non-sensitive personal data that becomes sensitive in an AIG-NCII image. Rather, what is sensitive is a technologically-produced representation of the self made possible using a small amount of personal data (e.g., a photograph of the subject) and a large amount of other people's data (used to train the model that generated the AIG-NCII). While technical work focusing on detecting sensitive parts of images [85] is valuable and should be continued, protecting representational privacy requires holistic considerations beyond just identifying and redacting sensitive image regions.

Legal scholars have already begun to wrestle with this issue, highlighting that existing regulation on privacy may not be wholly sufficient to protect sexual autonomy [18]. Citron proposes the recognition of sexual privacy — "the behaviors, expectations, and choices that manage access to and information about the human body, sex, sexuality, gender, and intimate activities" [18] — to provide more holistic protections for subjects of intimate images. What would a similar reformulation from data privacy to representational privacy mean for the technical security and privacy community? Answering this question will require translating notions of self-representation and consent into technical constraints that can govern systems.

6 Conclusion

Public familiarity with AIG-NCII is still low [86]. As more of it is produced [34] and it becomes easier to produce (e.g., through commercial text-to-video products or "nudify" apps [27, 61]), technological acceptance may increase and attitudes may change [40]. Continued work is needed to track and understand the development of technology for creating and sharing AIG-NCII as well as the attitudes around it. Our study contributes towards the understanding of attitudes towards non-consensual deepfakes across contexts, including AIG-NCII, providing insight into the rationales behind people's attitudes as well as the connections between gender, consent, genuine intimate imagery and these attitudes. Addressing AIG-NCII media requires a multifaceted response blending social science work on norms, legal scholarship, and socio-technical research to detect and prevent creation, sharing and viewing of harmful synthetic media.

Acknowledgments

References

[1] Executive Order No. 14110 - Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, 2023.

[2] Henry Ajder, Giorgio Patrini, Francesco Cavalli, and Laurence Cullen. The State of Deepfakes: Landscape, Threats, and Impact. Technical report, Deeptrace Labs, 2019.

[3] Haldun Akoglu. User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, 18(3):91–93, 2018.

[4] David Armstrong, Ann Gosling, John Weinman, and Theresa Marteau. The Place of Inter-Rater Reliability in Qualitative Research: An Empirical Study. *Journal of Sociology*, 31(2):597–606, August 1997.

[5] Roy Baker. *Defamation law and social attitudes: Ordinary unreasonable people*. Edward Elgar Publishing, 2011.

[6] Shaowen Bardzell and Jeffrey Bardzell. Towards a Feminist HCI Methodology: Social Science, Feminism, and HCI. In *Proc. CHI*, 2011.

[7] Samantha Bates. Revenge Porn and Mental Health: A Qualitative Analysis of the Mental Health Effects of Revenge Porn on Female Survivors. *Feminist Criminology*, 12(1):22–42, January 2017.

[8] Rasika Bhalerao, Vaughn Hamilton, Allison McDonald, Elissa M Redmiles, and Angelika Strohmayer. Ethical practices for security research with at-risk populations. In *2022 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 546–553. IEEE, 2022.

[9] Virginia Braun and Victoria Clarke. Can I use TA? Should I use TA? Should I not use TA? Comparing reflexive thematic analysis and other pattern-based qualitative analytic approaches. *Counselling and Psychotherapy Research*, 21(1):37–47, 2021.

[10] Virginia Braun and Victoria Clarke. One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology*, 18(3):328–352, 2021.

[11] Natalie Grace Brigham, Miranda Wei, Tadayoshi Kohno, and Elissa M. Redmiles. "Violation of my body:" Perceptions of AI-generated non-consensual (intimate) imagery. 2024. Available online at https://arxiv.org/abs/2406.05520.

[12] Roberto Caldelli, Leonardo Galteri, Irene Amerini, and Alberto Del Bimbo. Optical Flow based CNN for detection of unlearnt deepfake manipulations. *Pattern Recognition Letters*, 146:31–37, 2021.

[13] Nicholas Caporusso. Deepfakes for the Good: A Beneficial Application of Contentious Artificial Intelligence Technology. In Tareq Ahram, editor, *Advances in Artificial Intelligence, Software and Systems Engineering*, pages 235–241, Cham, 2021. Springer International Publishing.

[14] Pew Research Center. Assessing the Risks to Online Polls from Bogus Respondents. Technical report, Pew Research Center, February 18 2020.

[15] Robert Chesney and Danielle Citron. Deepfakes and the New Disinformation War: The Coming Age of Post-Truth Geopolitics. https://www.foreignaffairs.com/articles/world/2018-12-11/deepfakes-and-new-disinformation-war, January/February 2019. Accessed 29 January 2024.

[16] Lilie Chouliaraki. *The Spectatorship of Suffering*. Sage, 2006.

[17] Danielle K. Citron and Robert Chesney. Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *California Law Review*, 107:1753, 2019.

[18] Danielle Keats Citron. Sexual Privacy. *Yale Law Journal*, 128(7):1792–2121, May 2019.

[19] Valdemar Danry, Joanne Leong, Pat Pataranutaporn, Pulkit Tandon, Yimeng Liu, Roy Shilkrot, Parinya Punpongsanon, Tsachy Weissman, Pattie Maes, and Misha Sra. AI-Generated Characters: putting Deepfakes to Good Use. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–5, 2022.

[20] Audrey de Rancourt-Raymod and Nadia Smaili. The Unethical Use of Deepfakes. *Journal of Financial Crime*, 30(4):1066–1077, 2023.

[21] Rebecca A. Delfino. Pornographic Deepfakes: The Case for Federal Criminalization of Revenge Porn's Next Tragic Act. *Fordham Law Review*, 88:887, 2019.

[22] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton-Ferrer. The DeepFake Detection Challenge Dataset. *CoRR*, abs/2006.07397, 2020.

[23] Suzie Dunn. Technology-Facilitated Gender-Based Violence: An Overview. Technical report, Centre for International Governance Innovation, 2020.

[24] Suzie Dunn. Women, Not Politicians, Are Targeted Most Often by Deepfake Videos. https://www.cigionline.org/articles/women-not-politicians-are-targeted-most-often-deepfake-videos/?s=03, 2021.

[25] Asia Eaton, Holly Jacobs, and Yanet Ruvalcaba. Nationwide Online Study of Nonconsensual Porn Victimization and Perpetration. Technical report, Cyber Civil Rights Initiative, 2017.

[26] Victoria Elliott. Thinking about the Coding Process in Qualitative Data Analysis. *Qualitative Report*, 23:2850–2861, 11 2018.

[27] Kim Elsesser. Apps That Undress Women's Photos Surge In Popularity. What Could Go Wrong? https://www.forbes.com/sites/kimelsesser/2023/12/08/apps-that-undress-womens-photos-surge-in-popularity-what-could-go-wrong/?sh=73f783d923d3, December 2023. Accessed 10 February 2024.

[28] Pardis Emami-Naeini, Joseph Breda, Wei Dai, Tadayoshi Kohno, Kim Laine, Shwetak Patel, and Franziska Roesner. Understanding People's Concerns and Attitudes Toward Smart Cities. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery.

[29] Charles Ess. *Digital Media Ethics*. Polity Press, Cambridge, 2014.

[30] Hubert Etienne. The Future of Online Trust (and Why Deepfake Is Advancing It). *AI and Ethics*, 1(4):553–562, November 1 2021. ID: Etienne2021.

[31] Hany Farid. Creating, Using, Misusing, and Detecting Deep Fakes. *Journal of Online Trust and Safety*, 1(4), 2022.

[32] Dean Fido, Jaya Rao, and Craig A. Harper. Celebrity status, sex, and variation in psychopathy predicts judgements of and proclivity to generate and distribute deepfake pornography. *Computers in Human Behavior*, 129:107141, 2022.

[33] Asher Flynn, Elena Cama, Anastasia Powell, and Adrian J Scott. Victim-blaming and image-based sexual abuse. *Journal of Criminology*, 56(1):7–25, 2023.

[34] Asher Flynn, Anastasia Powell, Adrian J Scott, and Elena Cama. Deepfakes and Digitally Altered Imagery Abuse: A Cross-Country Exploration of an Emerging form of Image-Based Sexual Abuse. *The British Journal of Criminology*, 62(6):1341–1358, 12 2021.

[35] National Coalition for Sexual Freedom. Consent Counts. https://ncsfreedom.org/key-programs-2/consent-counts/#Consent_Counts_Statement. Accessed 9 February 2024.

[36] Dilrukshi Gamage, Piyush Ghasiya, Vamshi Bonagiri, Mark E. Whiting, and Kazutoshi Sasahara. Are Deepfakes Concerning? Analyzing Conversations of Deepfakes on Reddit and Exploring Societal Implications. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, pages 103:1–103:19, New York, NY, USA, 2022. Association for Computing Machinery.

[37] Anne Pechenik Gieseke. "The New Weapon of Choice": Law's Current Inability to Properly Address Deepfake Pornography. *Vanderbilt Law Review*, 73:1479, 2020.

[38] Roger Giner-Sorolla, Tom Kupfer, and John Sabo. What makes moral disgust special? an integrative functional review. In *Advances in experimental social psychology*, volume 57, pages 223–289. Elsevier, 2018.

[39] Jeffrey Gottfried. About three-quarters of Americans favor steps to restrict altered videos and images. Report, 2019. (2019).

[40] Andrina Granic. Technology Acceptance Model: a Literature Review from 1986 to 2013. *Universal Access in the Information Society*, 13(1):149–160, 2014.

[41] Jens Hainmueller, Dominik Hangartner, and Teppei Yamamoto. Validating vignette and conjoint survey experiments against real-world behavior. *Proceedings of the National Academy of Sciences*, 112(8):2395–2400, 2015.

[42] Andrew Gary Darwin Holmes. Researcher Positionality - A Consideration of Its Influence and Place in Qualitative Research - A New Researcher Guide. *International Journal of Education*, 8(4):1–10, 2020.

[43] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers, 2022.

[44] Antoinette Huber. 'A shadow of me old self': The impact of -based sexual abuse in a digital society. *International Review of Victimology*, 29(2):199–216, 2023.

[45] Terry P. Humphreys and Mélanie M. Brousseau. The Sexual Consent Scale–Revised: Development, Reliability, and Preliminary Validity. *The Journal of Sex Research*, 47(5):420–428, 2010. PMID: 19685367.

[46] Lilly Irani, Janet Vertesi, Paul Dourish, Kavita Philip, and Rebecca E Grinter. Postcolonial computing: a lens on design and development. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1311–1320, 2010.

[47] Farzaneh Karegar. *Towards Improving Transparency, Intervenability, and Consent in HCI*. PhD thesis, Karlstad University Press, 2018.

[48] Tsachi Keren-Paz. *10: The Power of Property: Strict Liability for Viewing NCII*, chapter 10, pages 175–193. Bristol University Press, 2023.

[49] Jan Kietzmann, Linda W. Lee, Ian P. McCarthy, and Tim C. Kietzmann. Deepfakes: Trick or treat? *Business Horizons*, 63(2):135–146, 2020.

[50] Tadayoshi Kohno, Yasemin Acar, and Wulf Loh. Ethical frameworks and computer security trolley problems: Foundations for conversations. In *USENIX Security*, 2023.

[51] Matthew B. Kugler and Carly Pace. Deepfake Privacy: Attitudes and Regulation. *Nw. UL Rev.*, 116:611, 2021.

[52] Minghui Li and Yan Wan. Norms or fun? The influence of ethical concerns and perceived enjoyment on the regulation of deepfake information. *Internet Research: Electronic Networking Applications and Policy*, 33(5):1750–1773, 2023.

[53] Emanuel Maiberg and Samantha Cole. AI-Generated Taylor Swift Porn Went Viral on Twitter. Here's How It Got There. https://www.404media.co/ai-generated-taylor-swift-porn-twitter/, January 2024. Accessed 29 January 2024.

[54] Nikola Marangunić and Andrina Granić. Technology acceptance model: a literature review from 1986 to 2013. *Universal access in the information society*, 14:81–95, 2015.

[55] Kirsten E. Martin. Diminished or Just Different? A Factorial Vignette Study of Privacy as a Social Contract. *Journal of Business Ethics*, 111(4):519–539, 2012.

[56] Clare McGlynn, Kelly Johnson, Erika Rackley, Nicola Henry, Nicola Gavey, Asher Flynn, and Anastasia Powell. 'It's Torture for the Soul': The Harms of Image-Based Sexual Abuse. *Social & Legal Studies*, 30(4):541–562, 2021.

[57] Clare McGlynn and Erika Rackley. Image-Based Sexual Abuse. *Oxford Journal of Legal Studies*, 37(3):534–561, 2017.

[58] Clare McGlynn, Erika Rackley, and Ruth Houghton. Beyond 'Revenge Porn': The Continuum of Image-Based Sexual Abuse. *Feminist Legal Studies*, 25:25–46, 2017.

[59] Annelise Mennicke, Jessi Bowling, Jennifer Gromer, and Caitlin Ryan. Factors Associated With and Barriers to Disclosure of a Sexual Assault to Formal On-Campus Resources Among College Students. *Violence Against Women*, 27(2):255–273, 2021.

[60] Janice M. Morse. Perfectly Healthy, But Dead: The Myth of Inter-Rater Reliability. *Qualitative Health Research*, 7(4):445–447, November 1997.

[61] Margi Murphy. 'Nudify' apps that use AI to undress women in photos are soaring in popularity. https://fortune.com/2023/12/08/nudify-apps-use-ai-popularity-deepfakes/, December 2023. Accessed 10 February 2024.

[62] Pardis Emami Naeini, Sruti Bhagavatula, Hana Habib, Martin Degeling, Lujo Bauer, Lorrie Faith Cranor, and Norman Sadeh. Privacy Expectations and Preferences in an IoT World. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*, pages 399–412, Santa Clara, CA, July 2017. USENIX Association.

[63] Stuart Napshin, Jomon Paul, and Justin Cochran. Individual Responsibility Around Deepfakes: It's No Laughing Matter. *Cyberpsychology, Behavior, and Social Networking*, 2024.

[64] Yuval Nirkin, Yosi Keller, and Tal Hassner. FSGAN: Subject Agnostic Face Swapping and Reenactment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[65] Helen Nissenbaum. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press, Stanford, CA, 2010.

[66] Lindsay M. Orchowski, Amy S. Untied, and Christine A. Gidycz. Social Reactions to Disclosure of Sexual Victimization and Adjustment Among Survivors of Sexual Assault. *Journal of Interpersonal Violence*, 28(10):2005–2023, 2013.

[67] Eyal Peer, David Rothschild, Andrew Gordon, Zak Evernden, and Ekaterina Damer. Data Quality of Platforms and Panels for Online Behavioral Research. *Behavior Research Methods*, 54(4):1643–1662, 08 2022. Published on August 1, 2022.

[68] Stephen R. Porter, Michael E. Whitcomb, and William H. Weitzer. Multiple Surveys of Students and Survey Fatigue. *New Directions for Institutional Research*, 121:63–73, 2004.

[69] Jeremy Prichard, Richard Wortley, Paul A Watters, Caroline Spiranovic, Charlotte Hunn, and Tony Krone. Effects of automated messages on internet users attempting to access "barely legal" pornography. *Sexual Abuse*, 34(1):106–124, 2022.

[70] Prolific. Why Prolific? https://www.prolific.co/prolific-vs-mturk/, 2020. Accessed 29 January 2024.

[71] Lisa R Pruitt. Her own good name: Two centuries of talk about chastity. *Md. L. Rev.*, 63:401, 2004.

[72] Albert Pumarola, Antonio Agudo, Adrià Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically aware facial animation from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 818–833, 2018.

[73] Reddit. Never Post Intimate or Sexually Explicit Media of Someone Without Their Consent. https://www.reddithelp.com/hc/en-us/articles/360043513411. Accessed 29 January 2024.

[74] Elissa M. Redmiles, Mia M. Bennett, and Tadayoshi Kohno. Power in computer security and privacy: A critical lens. *IEEE Security & Privacy Magazine*, March/April 2023.

[75] Elissa M Redmiles, Sean Kross, and Michelle L Mazurek. How Well Do My Results Generalize? Comparing Security and Privacy Survey Results from MTurk, Web, and Telephone Samples. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 1326–1343. IEEE, 2019.

[76] Nataniel Ruiz, Sarah Adel Bargal, Cihang Xie, and Stan Sclaroff. Practical Disruption of Image Translation Deepfake Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):14478–14486, Jun. 2023.

[77] Yanet Ruvalcaba and Asia A Eaton. Nonconsensual Pornography among US Adults: A Sexual Scripts Framework on Victimization, Perpetration, and Health Correlates for Women and Men. *Psychology of Violence*, 10(1):68, 2020.

[78] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-A-Video: Text-to-Video Generation without Text-Video Data, 2022.

[79] Gerald R Smith. Of malice and men: The law of defamation. *Val. UL Rev.*, 27:39, 1992.

[80] StopNCII. Frequently Asked Questions. Website, Accessed 2024.

[81] Daniel Story and Ryan Jenkins. Deepfake Pornography and the Ethics of Non-Veridical Representations. *Philosophy & Technology*, 36(3):56, August 26 2023.

[82] Yolande Strengers, Jathan Sadowski, Zhuying Li, Anna Shimshak, and Florian 'Floyd'Mueller. What can HCI learn from sexual consent? A feminist process of embodied consent for interactions with emerging technologies. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2021.

[83] Grace Tatter. Consent at Every Age. https://www.gse.harvard.edu/ideas/usable-knowledge/18/12/consent-every-age, December 2018. Accessed 9 February 2024.

[84] Brian Timmerman, Pulak Mehta, Progga Deb, Kevin Gallagher, Brendan Dolan-Gavitt, Siddharth Garg, and Rachel Greenstadt. Studying the Online Deepfake Community. *Journal of Online Trust and Safety*, 2(1), Sep. 2023.

[85] Ashwini Tonge and Cornelia Caragea. Image Privacy Prediction Using Deep Neural Networks. *ACM Trans. Web*, 14(2), apr 2020.

[86] Rebecca Umbach, Nicola Henry, Gemma Faye Beard, and Colleen M. Berryessa. Non-consensual synthetic intimate imagery: Prevalence, attitudes, and knowledge in 10 countries. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.

[87] Nadia de Vries. "Porsche Girl": When a Dead Body Becomes a Meme. *MIT Case Studies in Social and Ethical Responsibilities of Computing*, (Summer 2022), aug 26 2022. https://mit-serc.pubpub.org/pub/porsche-girl.

[88] Miranda Wei, Pardis Emami-Naeini, Franziska Roesner, and Tadayoshi Kohno. Skilled or Gullible? Gender

Stereotypes Related to Computer Security and Privacy. In *IEEE Symposium on Security and Privacy*, 2023.

[89] Mika Westerlund. The Emergence of Deepfake Technology: A Review. *Technology Innovation Management Review*, 9(11):9–16, 2019.

[90] Robin Whittemore, Susan K. Chase, and Carol Lynn Mandle. Validity in Qualitative Research. *Qualitative Health Research*, 11(4):522–537, July 2001.

[91] David Gray Widder, Dawn Nafus, Laura Dabbish, and James Herbsleb. Limits and Possibilities for "Ethical AI" in Open Source: A Study of Deepfakes. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2035–2046, New York, NY, USA, 2022. Association for Computing Machinery.

[92] Kaylee Williams. Exploring Legal Approaches to Regulating Nonconsensual Deepfake Pornography, 2023.

[93] Rhiannon Williams. Text-to-image AI models can be tricked into generating disturbing images. *MIT Technology Review*, 2023.

[94] Andrea L. Wirtz, Tonia C. Poteat, Mannat Malik, and Nancy Glass. Gender-Based Violence Against Transgender People in the United States: A Call for Research and Programming. *Trauma, Violence, & Abuse*, 21(2):227–241, 2020.

[95] Pamela J Wisniewski and Xinru Page. Privacy theories and frameworks. In *Modern Socio-Technical Perspectives on Privacy*, pages 15–41. Springer International Publishing Cham, 2022.

[96] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7623–7633, October 2023.

[97] Zhiliang Xu, Zhibin Hong, Changxing Ding, Zhen Zhu, Junyu Han, Jingtuo Liu, and Errui Ding. Mobilefaceswap: A lightweight framework for video face swapping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2973–2981, 2022.

[98] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-Attentional Deepfake Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2185–2194, June 2021.

[99] Douglas Zytko, Jane Im, and Jonathan Zong. Consent: A Research and Design Lens for Human-Computer Interaction. In *Companion Publication of the 2022 Conference on Computer Supported Cooperative Work and Social Computing*, pages 205–208, 2022.

[100] Carl Öhman. Introducing the Pervert's Dilemma: A Contribution to the Critique of Deepfake Pornography. *Ethics and Information Technology*, 22(2):133–140, June 2020.

A   Participant demographics

Participants' gender, age, and political orientation is presented in Table A.

B   Media action regression results

Results for the regression are presented in Table 6.

C   Qualitative Codebook

The codebooks from qualitatively analyzing explanations for why the creation of the synthetic video in each vignettes is either acceptable or unacceptable. Codes were not mutually exclusive.

Rationales for acceptability:

- **No Harm**: Will not cause harm
- **Relationship**: Trust in an intimate partner
- **Indifference**: No impact; 'I don't care'
- **Compliment**: Indicates attraction
- **Fantasy**: Indulges fantasy
- **Pro-Tech**: Technology and AI are interesting

Rationales for unacceptability:

- **Consent**: Absence of consent or permission
- **Awareness**: Lack of awareness about video's creation and existence
- **Dislike**: Elicits negative feelings; The video is 'weird,' 'creepy,' 'disgusting,' 'uncomfortable,' etc.
- **Harm**: Creates or could create harm
- **Ethics**: Violation of ethics, morality, or law; The video is 'wrong'
- **Privacy**: Violation of privacy
- **Fake**: Fake nature, inauthentic
- **Stranger**: Created by a stranger
- **Relationship**: Violation of trust in an intimate partner

Gender		Age		Political Orientation	
		18-24	17.8%		
Woman	49.5%	25-34	33.0%	Democrat	48.6%
Man	47.6%	35-44	24.4%	Republican	16.2%
Non-binary	1.9%	45-54	13.3%	Leans Democrat	18.4%
Agender	0.6%	55-64	7.9%	Leans Republican	8.9%
Prefer not to say	0.3%	65+	2.9%	Refuse to answer	7.9%
		Prefer not to say	0.6%		

Table 5: Breakdown of participant demographics by gender, age, and political orientation.

		OR; Confidence Interval
Intercepts	Totally unacceptable | Somewhat unacceptable	2.42; [1.89, 3.1]***
	Somewhat unacceptable | Neutral	7.41; [5.73, 9.59]***
	Neutral | Somewhat acceptable	28.59; [21.65, 37.76]***
	Somewhat acceptable | Totally acceptable	89.53;[65.78, 121.85]***
Content Action	private_sharing	0.47; [0.37, 0.58]***
	public_sharing	0.26; [0.21, 0.33]***
	resharing	0.42; [0.33, 0.52]***
	seeking_out	5.43; [4.45, 6.62]***

Table 6: Results from a single regression exploring the relationship between acceptability (first row, intercepts) and action being preformed with the synthetic media (second row, content action). Reference level of content action is creation. Significance of OR: $p < 0.001 =$ ***.

D   Additional Models

Regression analyses conducted with gender categorized into 'men' and 'women,' rather than 'men' and 'marginalized genders.' Eight participants who identified outside of the gender binary or did not disclose their gender were excluded from these analyses. For the results with gender bucketed into 'men' and 'women,' Table 7 corresponds to Table 2, Table 9 to Table 3, and Table 8 to Table 4.

		creation	private_sharing	public_sharing	resharing	seeking_out
Intercepts	Totally unacceptable | Somewhat unacceptable	5.86 [0.47, 73.58]	1.48 [0.07, 29.23]	22.30 [0.64, 779.31]	1.36 [0.05, 24.06]	0.03* [0, 1]
	Somewhat unacceptable | Neutral	34.07** [2.66, 436.98]	7.09 [0.35, 141.55]	91.48* [2.54, 3297.44]	6.03 [0.24, 152.77]	0.18 [0.01, 5.37]
	Neutral | Somewhat acceptable	114.61*** [8.77, 1497.84]	22.42* [1.11, 451.16]	216.15** [5.89, 7933.25]	19.72 [0.77, 503.89]	6.08 [0.21, 174.27]
	Somewhat acceptable | Totally acceptable	426.21*** [31.71, 5728.35]	101.08** [4.89, 2089.5]	1104.00*** [28.58, 42643.48]	91.20** [3.47, 2398.75]	28.93* [1, 835.39]
Controlled IVs	creator (Intimate partner)	3.29*** [2.25, 4.79]	1.71* [1.13, 2.58]	1.45 [0.89, 2.37]	1.04 [0.67, 1.6]	1.13 [0.81, 1.57]
	action (Sport)	12.96*** [7.69, 21.85]	33.43*** [16.13, 69.26]	64.11*** [22.29, 184.37]	31.32*** [14.57, 67.29]	7.22*** [4.68, 11.15]
	action (Saying something)	5.48*** [3.29, 9.14]	10.58*** [5.24, 21.34]	19.33*** [6.92, 53.97]	12.26*** [5.79, 25.95]	3.39*** [2.19, 5.23]
	intent (Entertainment)	19.27*** [11.17, 33.24]	12.04*** [6.85, 21.17]	10.89*** [5.59, 21.19]	5.80*** [3.132, 10.15]	4.83*** [3.16, 7.4]
	intent (Sexual pleasure)	7.51*** [4.45, 12.68]	1.42 [0.81, 2.5]	1.17 [0.59, 2.32]	0.92 [0.52, 1.64]	1.29 [0.86, 1.92]
Uncontrolled IVs	Gender (Man)	2.35** [1.39, 3.99]	2.06* [1.18, 3.61]	1.64 [0.82, 3.27]	1.38 [0.73, 2.63]	1.56 [0.78, 3.12]
	GII & NDII attitudes (Unacceptable)	0.21* [0.05, 0.85]	0.08** [0.01, 0.4]	0.09** [0.02, 0.43]	0.01*** [0, 0.05]	0.01*** [0.01, 0.03]
	SCS-R2	0.53*** [0.39, 0.72]	0.56*** [0.4, 0.77]	0.64* [0.43, 0.96]	0.76 [0.52, 1.11]	0.73 [0.49, 1.11]
	SCS-R4	1.10 [0.85, 1.43]	1.14 [0.87, 1.5]	1.25 [0.89, 1.76]	1.33 [0.97, 1.83]	1.17 [0.84, 1.65]

Table 7: Results from regressions exploring the relationship between scenario acceptability (first row, intercepts), contextual factors (second row, controlled IVs), and personal factors (third row, uncontrolled IVs). Each column represents the output of one regression model. Numeric cells list the odds ratio (OR) and the 95% confidence interval. Reference levels: creator (stranger), action (sexual act), intent (harm), gender (woman), GII & NDII attitudes (acceptable). Significance of OR: $p < 0.05 =$ * , $p < 0.01 =$ ** , and $p < 0.001 =$ *** .

Table 8

		OR; Confidence Interval
Intercepts	Totally unacceptable | Somewhat unacceptable	5.83; [0.36, 80.73]
	Somewhat unacceptable | Neutral	33.60; [2.2, 513.89]*
	Neutral | Somewhat acceptable	118.44; [7.63, 1838.84]***
	Somewhat acceptable | Totally acceptable	462.15; [29, 7364.66]***
Controlled IVs	creator (Intimate partner)	1.76; [1.03, 3.01]*
	action (Sport)	16.03;[7.31, 35.16]***
	action (Saying something)	10.95; [4.92, 24.37]***
	intent (Entertainment)	20.64; [9.49, 44.9]***
	intent (Sexual pleasure)	4.90; [2.29, 10.47]***
Uncontrolled IVs	Gender (man)	1.44; [0.39, 5.27]
	GII & NDII attitudes (Unacceptable)	0.2; [0.05, 0.85]*
	SCS-R2	0.52; [0.38, 0.72]***
	SCS-R4	1.13; [0.86, 1.47]
Interaction Terms	action (Sport) & Gender (Man)	0.84; [0.32, 2.23]
	action (Saying something) & Gender (Man)	0.31; [0.11, 0.86]*
	intent (Entertainment) & Gender (Man)	0.97; [0.37, 2.59]
	intent (Sexual pleasure) & Gender (Man)	2.38; [0.86, 6.53]
	creator (Intimate partner) & Gender (Man)	3.58; [1.7, 7.56]***

Table 8: Results from a single regression exploring the relationship between scenario acceptability for creation (first row, intercepts), contextual factors (second row, controlled IVs), personal factors (third row, uncontrolled IVs), and interactions between gender and contextual factors (third row, interaction terms). Reference levels: creator (stranger), action (sexual act), intent (harm), gender (woman), GII & NDII attitudes (acceptable). Significance of OR: $p < 0.05$ = *, $p < 0.01$ = **, and $p < 0.001$ = ***.

Table 9

		OR; Confidence Interval
Intercepts	Totally unacceptable | Somewhat unacceptable	8.90; [0.49, 163.28]
	Somewhat unacceptable | Neutral	57.23; [3.07, 1067.96]**
	Neutral | Somewhat acceptable	204.33; [10.78, 3872.9]***
	Somewhat acceptable | Totally acceptable	784.29; [40.39, 15230]***
Controlled IVs	creator (Intimate partner)	1.46; [0.66, 3.26]
	action (Sport)	47.55; [11.08, 204.08]***
	action (Saying something)	9.96; [2.25, 44.08]**
	intent (Entertainment)	13.88; [3, 64.17]***
	intent (Sexual pleasure)	21.13; [4.57, 97.62]***
Uncontrolled IVs	Gender (man)	2.54; [1.46, 4.41]***
	GII & NDII attitudes (Unacceptable)	0.19; [0.04, 0.81]*
	SCS-R2	0.51; [0.37, 0.71]***
	SCS-R4	1.13; [0.86, 1.48]
Interaction Terms	creator (Intimate partner) & intent (Entertainment)	2.62; [0.98, 7.01]
	creator (Intimate partner) & intent (Sexual pleasure)	3.64; [1.33, 9.96]**
	action (Sport) & intent (Entertainment)	0.72; [0.14, 3.54]
	action (Saying something) & intent (Entertainment)	1.70; [0.33, 8.82]
	action (Sport) & intent (Sexual pleasure)	0.08; [0.02, 0.4]**
	action (Saying something) & intent (Sexual pleasure)	0.19; [0.04, 1]*

Table 9: Results from a single regression exploring the relationship between the acceptability of creation (first row, intercepts), contextual factors (second row, controlled IVs), personal factors (third row, uncontrolled IVs), and interactions between intent and creator or action (fourth row, interaction terms). Reference levels: creator (stranger), action (sexual act), intent (harm), gender (woman), GII & NDII attitudes (acceptable). Significance of OR: $p < 0.05$ = *, $p < 0.01$ = **, and $p < 0.001$ = ***.

What Drives SMiShing Susceptibility? A U.S. Interview Study of How and Why Mobile Phone Users Judge Text Messages to be Real or Fake

Sarah Tabassum
University of North Carolina at Charlotte

Cori Faklaris
University of North Carolina at Charlotte

Heather Richter Lipford
University of North Carolina at Charlotte

Abstract

In today's digital world, SMS phishing, also known as SMiShing, poses a serious threat to mobile users. However, it is unclear whether existing research on phishing can be applied to SMiShing. Our study aims to fill this gap by conducting interviews with 29 mobile phone users in a major southeastern U.S. city. We collected data on participants' experiences with suspicious SMS, understanding the cues they pay attention to, how they verify and report such messages, and the role of prior training in distinguishing real messages from scams. We also collected data on how specific details and context make a legitimate SMS seem genuine. Our findings indicate that participants focus more on the content, format, and links in SMS rather than the sender's short code, phone number, or email address. We suggest design changes to enhance user awareness and resilience against SMS phishing. This research provides practical knowledge to mitigate cyber threats linked to SMiShing. To the best of our knowledge, this is the first interview study on SMiShing susceptibility.

1 Introduction

With the continuous global surge in mobile phone adoption, as of January 2024, approximately two-thirds of the world's population, totaling 5.44 billion people, are actively using mobile phones [32]. Integral to every mobile phone is the Short Message Service (SMS) feature, which, according to Keepnet Labs, has become a prevalent medium for phishing attacks, especially since the onset of the COVID-19 pandemic. In 2020, SMS phishing or smishing attacks saw a staggering 328%

increase, with 76% of businesses being targeted during that period [37]. The prominence of text message fraud is further underscored by data from the US Federal Trade Commission (FTC) in 2023, where text message fraud ranked among the top three methods employed by scammers, alongside emails and phone calls [5]. The financial implications of these scams were substantial, with a total loss of $10 billion reported in 2023, of which $372 million resulted solely from fraud text messages [5].

Understanding the reasons behind people falling victim to such scams via text messages is crucial. While extensive research exists on email phishing, its email-based cousin [9, 17, 30, 41, 46, 47, 49], the effectiveness of these techniques in SMShing remains unexplored. Both exploit trust and urgency for deception, but SMS communication's unique features such as shorter length, limited information, and immediacy create a distinct playing field [26, 39]. Traditional phishing research findings, built on email analysis and user behavior, may not directly translate to this mobile-based phishing.

Insights from the 2023 Databook by the US Federal Trade Commission are that younger individuals in the US are more susceptible to these fraud SMS scams, reporting their experiences, while older individuals tend to incur higher financial losses [5]. However, little to no published peer-reviewed research explores why people fall victim to SMiShing or the cues they use to identify legitimate and fraudulent SMS. To address this gap, our study begins with asking, *How do individual participants perceive the credibility of SMS messages and make trust decisions?* And, *What individual and design factors seem important?* To answer these inquiries, we employed an interview-based approach with mobile phone users in a major U.S. city, inspired by Jakobsson's work on phishing email cues [30]. We interviewed 29 participants, focusing on their mobile phone usage, experiences with suspicious and fraudulent SMS, cues they look for to distinguish legitimate from fraudulent SMS, effect of cybersecurity training, verification practices, and reporting behavior. During the interviews, participants were also asked to identify fraudulent and legitimate SMS from examples we provided, based on

specific cues. We employed inductive methods to analyze the interviews.

Our findings revealed that while determining the legitimacy of an SMS, they prioritized cues such as contents with links, misspelled and out-of-context messages as suspicious indicators. For legitimate text messages, they looked for personalized information, a familiar context, known senders, and an official format. Those who had received some cybersecurity training demonstrated better judgment than those without training. Interestingly, most individuals did not report suspicious messages; instead, they tended to ignore them. Our study suggests a need for increased awareness, the implementation of SMS spam filters on iOS similar to Android [36], and an improved user interface for reporting fraudulent messages.

As of our knowledge, this study represents the first qualitative exploration into SMiShing. Our paper contributes the following:

- An enhanced understanding of individuals' real-life experiences with SMS phishing/fraud SMS.

- Insights into the cues that people use to distinguish between legitimate and fraudulent SMS.

- Design suggestions for telecommunication and mobile companies based on user data.

2 Background and Related Work

Phishing, a persistent cybersecurity threat, has undergone significant evolution since its emergence in the mid-'90s [23]. Initially recognized as a serious concern, service providers began responding with intensified efforts, deploying technical, educational, and legal interventions [30]. Despite these countermeasures, the Anti-Phishing Working Group (APWG) reported a staggering 1,286,208 phishing attacks in the second quarter of 2023, with the financial sector being the primary target, accounting for 23.5% of all attacks [10]. A notable evolution in the phishing landscape was the rise of SMS phishing, commonly known as SMiShing [43]. This variant gained prominence as attackers exploit text messages to deceive users [14]. SMiShing typically involves the dissemination of fraudulent links in text messages, leading unsuspecting victims to forms designed to either extract sensitive information or download malicious content [14, 43].

2.1 Phishing Attacks

Phishing is primarily employs fraudulent emails to impersonate legitimate entities and solicit sensitive information from users [11, 13, 28, 30]. This is recognized as one of the most common and extensively studied cyberattacks [1]. These emails often contain malicious links or attachments, redirecting users to fake websites or initiating the download of malware onto their devices [9, 28]. The motives behind phishing

attacks vary, ranging from stealing money and identities to credentials or intellectual property.

Efforts to prevent or detect phishing attacks have led to various research approaches. Hong suggests strategies such as "making things invisible," utilizing machine learning on the backend to classify and filter out phishing attempts, developing improved user interfaces, and providing effective training [18, 20]. Numerous studies have explored factors influencing user susceptibility to phishing attacks, including email design, message content, situational context, and user characteristics [9, 16, 19–21, 31, 41, 46].

Among the studies that focused on visual cues for distinguishing legitimate websites/links, it is worth mentioning the research conducted by Alsharnouby et al. and Petelka et al. Alsharnouby et al.'s investigation explored users' ability to identify legitimate websites by capturing their attention [9]. The study found that users could successfully identify only 53% of phishing websites. Moreover, the study revealed that users typically allocate minimal time to inspecting security indicators and mainly focus on the website content during their assessments [9]. The study by Petelka et al. examined the impact of relocating phishing warnings close to suspicious links in emails [41]. Their findings showed that link-focused phishing warnings significantly reduced click-through rates compared to email banner warnings [41].

Sheng et al. explored demographic vulnerability, revealing the heightened susceptibility of young females to phishing attacks [46]. Their findings underscored that women exhibited greater vulnerability than men, and participants aged 18 to 25 were particularly susceptible due to disparities in computer and web expertise. Educational materials were identified as effective in reducing participants' willingness to provide information on fake webpages, with a marginal decrease in users' inclination to click on legitimate links. In 2007, Jakobsson et al. conducted a study on user reactions to various "trust indicators" in both authentic and phishing stimuli, offering insights into what renders phishing emails and web pages authentic. This research not only guided the design of legitimate material to mitigate risks but also examined factors influencing consumers' perception of legitimate content as dubious, with potential implications for online advertising [30].

Motivated by insights derived from studies on phishing, particularly the works of Jakobsson, Sheng and Alsharnouby [9, 30, 46], our study concentrates on SMiShing. The aim is to adapt and expand upon the understanding of user vulnerabilities in the context of SMS phishing attacks.

2.2 SMiShing Attacks

The term SMiShing is derived from the fusion of SMS, which stands for Short Message Service - the technology underpinning text messages - and phishing [2, 43]. The use of SMS for malicious purposes, termed SMiShing, has been documented since the early 2000s [23]. SMiShing constitutes a social engi-

neering attack that leverages deceptive mobile text messages to deceive individuals into downloading malware, disclosing sensitive information, or transferring funds to cybercriminals. This form of cybercrime has gained increasing prevalence and sophistication over the years [4]. In 2022, 76% of organizations in the U.S. encountered SMiShing attacks [23].

Despite the increasing prevalence of SMiShing, there has been limited academic exploration of its vulnerabilities. Some studies focus on characterizing modern SMS phishing attacks, exemplified by Nahapetyan's work [40]. This research utilized public SMS gateways to capture 67,991 phishing messages over a period of 396 days, providing valuable insights into SMS phishing trends and the clustering of phishing operations. Moreover, Jakobsson's insightful article addresses the use of two-factor "inauthentication" and the growing prominence of SMS phishing attacks related to two-factor authentication [29]. In a study by Rahman et al., involving 10,000 participants exposed to various smishing attacks, they found that personalized or spoofed messages heightened the perceived legitimacy and urgency for users to respond [42]. In a recent survey study on SMiShing susceptibility, findings indicated that the younger population was more vulnerable to such attacks [22]. However, little is known about how well the results from phishing studies apply in this new context. Consequently, there is a need for further research to comprehend SMiShing vulnerabilities. Our work seeks to fill this research gap by gaining deeper insights on SMiShing attacks. Through our study, we aim to offer insights into the dynamics of SMiShing attacks and enhance the understanding of user vulnerabilities in this context.

3 Methodology

To understand the participants' thought processes and personal experiences relevant to our research questions, we conducted in-person interviews. This section discusses the recruitment process, participant demographics, details about the interview sessions, and the data analysis process employed in our study.

3.1 Recruitment

In this interview study, we interviewed 29 participants (16 Females, 13 Males). We promoted recruitment through university research announcements, flyers distributed in a major southeastern U.S. city, and advertisements placed on Craigslist, Facebook, and LinkedIn. These recruitment materials were designed to engage individuals who have encountered or are open to discussing fraudulent messages. Prospective participants were required to complete a brief eligibility survey to determine their eligibility for the interview. From the pool of eligible participants, we aimed to achieve diversity in terms of education/job status and gender. The selection

process involved evaluating factors such as gender and educational background. Individuals who met the criteria for the final interview were contacted via email and provided with a consent form. Participants who gave their consent for the interview were subsequently provided with information about the interview's available time slots and locations.

3.2 Participants

We have recruited 29 individuals who regularly use mobile phones, are 18 years or older, and reside in the metro area, making them available for in-person interviews. The participants consist of 16 females (55.2%) and 13 males (44.8%), spanning various age groups: 15 in the 18-24 range, 5 in the 25-34 range, 5 in the 35-44 range, 2 in 45-54 range and 2 who are 55 or older. Prior research indicates that the perception of cyber security risks and attention to trust indicators may differ based on age [25, 35]. In this study, we intentionally recruited people from different age groups to explore potential differences in their thought processes [50]. The participants also represent diverse professional backgrounds, including students, full-time employees, part-time employees, unemployed individuals, and self-employed individuals. Their professional backgrounds cover a wide spectrum, including computer science (CS), engineering (Eng.), business/management (BM), biology (Bio.), humanities (Hum.), education (Edu.), and even entertainment (Entr.).

Table 1 provides information about our participants on their age group (Age), gender (Gen.), mobile phone and carrier type (Mobile Set & Carrier), and occupation with the corresponding fields (Occupation). Furthermore, the table categorizes participants based on their professional status, including students ("Stu."), full-time employees ("FTE"), part-time employees ("PTE"), self-employed individuals ("SE"), and those currently unemployed ("U").

All participants utilize smartphones, with a variety of brands such as Samsung (Sam.), iPhone (iPhn.), Motorola (Moto.), Google Pixel (Pxl.), and Wiko Phone (Wiko). Their mobile carriers include AT&T, T-Mobile (T-Mob.), Tracfone (Trac.), H2O, Verizon (Vrzn), and Assurance (Asr.). Among our participants, 58.6% reported using their mobile phones for at least 21 to more than 30 hours in the past week at the time of the interview. During the interviews, participants were requested to bring their mobile phones to facilitate the review of text messages and the capture of screenshots if necessary.

3.3 Interview Sessions

During the interviews our team's researchers met with some participants at local coffee shops and others at the usability lab on campus. At the beginning of each session, the interviewer provided a brief introduction to the study's objectives and then asked for verbal consent to audio record the interview. Upon obtaining consent, the interviewer proceeded to

Table 1: Participant Demographics: Age group, gender, mobile phone and carrier information, and current occupation with corresponding field of study or profession are presented for all study participants

ID	Age	Gen.	Mobile Set & Carrier	Occupation
P1	25-34	M	Sam.(H2O)	Stu.(Civil Eng.)
P2	25-34	M	iPhn.(Vrzn)	Stu.(CS)
P3	55+	M	Moto.(T-Mob.)	FTE(Edu.)
P4	35-44	F	Sam.(AT&T)	SE(BM)
P5	35-44	F	iPhn.(T-Mob.)	FTE(Other)
P6	25-34	M	iPhn(T-Mob.)	Stu.(CS)
P7	25-34	M	Pxl.(AT&T)	Stu.(CS)
P8	18-24	F	Wiko(Asr.)	Stu.(Bio.)
P9	18-24	F	iPhn.(T-Mob.)	Stu.(CS)
P10	18-24	M	iPhn.(AT&T)	PTE(Other)
P11	18-24	F	iPhn.(AT&T)	U
P12	35-44	M	iPhn.(T-Mob.)	FTE.(Eng.)
P13	55+	F	iPhn.(AT&T)	FTE(BM)
P14	18-24	F	iPhn.(Vrzn)	Stu.(BM)
P15	18-24	F	iPhn.(Vrzn)	PTE(Other)
P16	18-24	F	Sam.(Vrzn)	Stu.(Hum.)
P17	18-24	F	iPhn.(Vrzn)	FTE(Edu.)
P18	25-34	M	iPhn(AT&T)	Stu.(BM)
P19	18-24	F	iPhn.(Trac.)	PTE(Other)
P20	45-54	F	iPhn.(Vrzn)	FTE(BM)
P21	45-54	M	iPhn.(Vrzn)	FTE(Edu.)
P22	18-24	M	iPhn.(Vrzn)	Stu.(BM)
P23	35-44	F	Sam.(T-Mob.)	FTE(Edu.)
P24	18-24	M	Sam.(Vrzn)	Stu.(EE)
P25	18-24	F	iPhn.(Vrzn)	Stu.(BM)
P26	18-24	M	Pxl.(AT&T)	Stu.(CS)
P27	35-44	F	Sam.(T-Mob.)	FTE(CS)
P28	18-24	M	iPhn.(Vrzn)	PTE(Entr.)
P29	18-24	F	iPhn.(T-Mob.)	Stu.(CS)

ask questions designed to address our research questions. The initial query focused on the participants' frequency of using texting apps, followed by questions about their mobile phone models and the mobile carriers they used. Subsequently, participants were invited to share their personal experiences with suspicious, fraud, and spam text messages. In this phase, participants were asked if they had any examples on their phones that they could share. Nearly all participants reported having multiple instances of suspicious and irritating messages. They were then asked to elaborate on why they considered those messages suspicious and were requested to share screenshots of the text messages. Next, each participant was presented with three pairs of legitimate and fraudulent text messages, chosen from a total of six pairs. The selection process was pseudo-random, with the interviewer counterbalancing the

pairs to ensure each participant was sufficiently exposed to a variety of messages. We instructed them to think-aloud so that we can understand their thought process. The provided examples, all six pairs, were determined through internal discussions with our research team and industry professionals. In the process of choosing SMS pair examples, we took into consideration the findings from the CSN Data Book 2023, which highlighted imposter scams as the #1 category among the top 10 fraud classifications, as reported by the U.S. Federal Trade Commission [5]. These scams involved the impersonation of bank authorities, government officials, and various services, including healthcare, online shopping transactions, and more. In our study, we concentrated on a diverse set of examples related to banks, credit cards, money transfers, online shopping, and package delivery. We included both iOS and Android messaging app interfaces for these scenarios. Out of the six pairs, four involved simulated bank-related text messages. The remaining two pairs represented real-life instances of both fraudulent and legitimate SMS. Figure 1 illustrates the four pairs of simulated bank-related SMS.

Figure 2 displays the remaining two pairs, taken from real-life instances involving credit/debit cards, package delivery, and online orders. Upon completion of the task, participants were asked about the initial aspects they noticed in a text message from an unfamiliar source and the elements that made a text appear suspicious or legitimate to them. We inquired separately about visual elements, icons, symbols, or colors they considered while assessing the credibility of an SMS. Additionally, participants were questioned about their preferred methods of verifying suspicious SMS messages, actions taken upon receiving such texts, any history of reporting such texts, and the outcomes of such reporting. We also explored their prior training in computer or cybersecurity and how it might aid them in efficiently identifying fraudulent SMS messages. Towards the end of the session, participants were asked if they had any expectations or suggestions that could assist them in recognizing malicious SMS more efficiently. Finally, we expressed gratitude to the participants for their valuable time. Additionally, we provided each participant with "Best Practices to Identify Fraudulent Text Messages" and expressed our appreciation by offering a $25 Amazon e-gift card for their participation. Each interview session ranged in duration from 35 to 56 minutes. The Institutional Review Board (IRB) at our university reviewed and approved our study, and we obtained informed consent from participants.

3.4 Data Analysis

As the interview sessions were conducted in person, we obtained participants' consent and recorded the audio for each session as a reference. Subsequently, we employed an automated transcription service to transcribe all recordings. The first author reviewed all transcripts to ensure alignment with the original recordings. Throughout the interview process, we
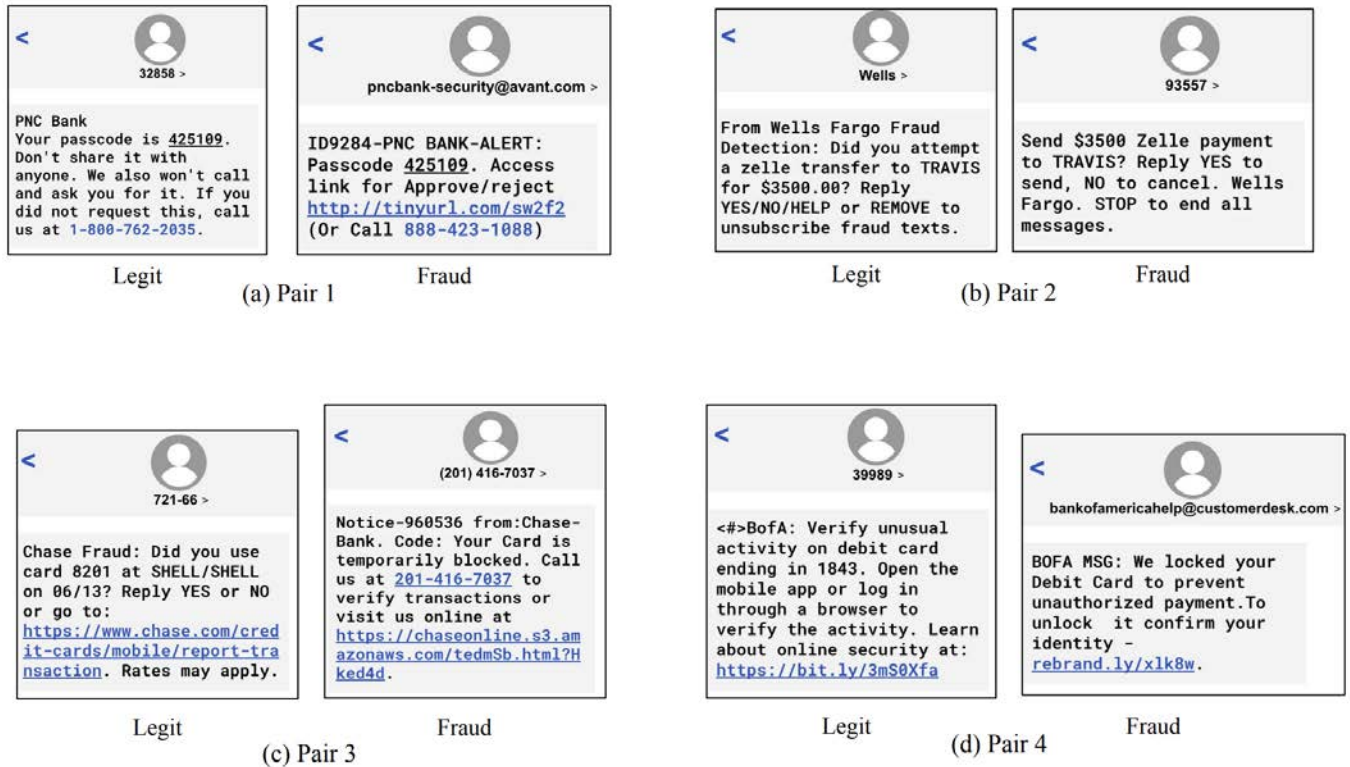
Figure 1: Visual representation of four pairs of simulated legitimate and fraudulent bank-related text messages presented to our participants. These text messages showcase deceptive tactics commonly employed in digital fraud.

systematically applied qualitative coding, enabling us to identify saturation points and adjust the interviews as interesting ideas or themes emerged. We analyzed our interview data using thematic analysis [12] and an inductive method [45], aligning with our research questions. We began by familiarizing ourselves with the data, then performed open coding [33] to segment it based on interview questions. We identified initial codes, explored similarities and differences, merged codes as needed, and organized them into themes, including cues for identifying suspicious and legitimate SMS messages. Research concluded when the codebook was completed, which is included in the appendix. In the coding phase, the first author completed coding for all transcripts. Inductive coding was then conducted across the entire dataset to develop a codebook. Afterward, both the first and second authors jointly reviewed the codes, resolving disagreements through discussion. Although the research team collaborated on code development and evaluation, the first author coded the entire dataset, eliminating the need for inter-rater reliability calculations [38].

4 Results

4.1 Mobile Phones, Carriers and Texting App Usage

Among the participants, 19 individuals (65.5%) were iPhone users with iOS. Android users showed diversity, with 6 participants using Samsung, 2 using Google Pixel, 1 using Wiko phone, and 1 using Motorola. Participants displayed varied choices in mobile carriers, with 11 users for Verizon, 8 users for T-Mobile, and 7 for AT&T. Additionally, there was 1 subscriber each for Assurance, H2O, and Tracfone.

All of our participants shared that they use texting apps on their phones on a daily basis. Participants found these apps very convenient for communicating with friends and family, as well as for planning activities. Many acknowledged regularly receiving suspicious or irritating SMS through these applications. Participants who underwent carrier switches generally reported no noticeable change in the frequency of spam SMS. However, it is noteworthy that one participant (P10) experienced an increase in spam calls and SMS after transitioning from T-Mobile to AT&T.

(a) Pair 5

(b) Pair 6
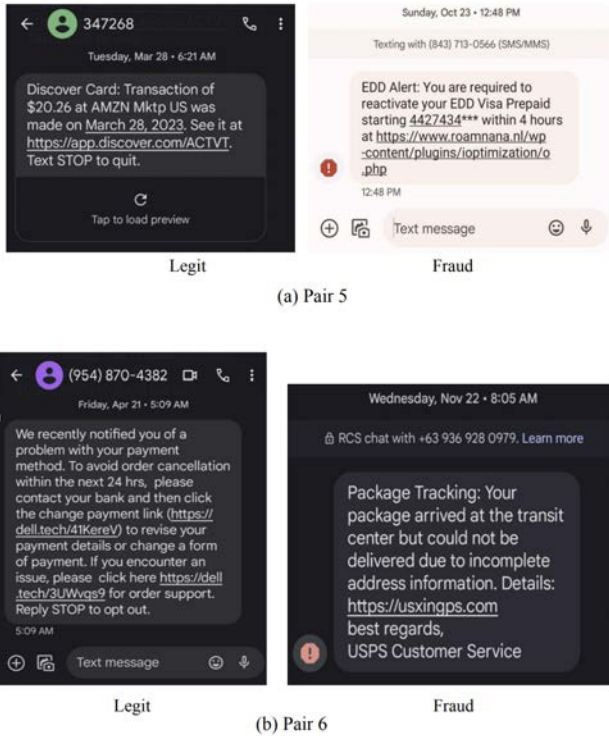
Figure 2: Two pairs of text messages, comprising both legitimate and fraudulent examples, gathered from real-life incidents

4.2 Personal Encounters with SMiShing

To gather insights into our participants' experiences with SMS phishing attacks, we queried, *"Have you ever received fraudulent or suspicious text messages on your phone, especially in the last 3 months?"* In response, all participants reported receiving such messages, with most encountering at least 1 or 2 per month, and some experiencing 3-4 weekly.

Subsequently, we requested participants to share specific examples from their phones. Out of 29 participants, 25 shared SMS examples. Analyzing these messages, we sought to understand the criteria participants used to identify fraudulent SMS and explore any discernible patterns. This investigation revealed varying types of fraud, with certain SMS categories prevailing over others. The following discussion presents the identified SMS types in order of prevalence.

4.2.1 Package Delivery Fraud

Among participants sharing suspicious SMS, 44% (11 out of 25) reported receiving messages purportedly from USPS or UPS. Notably, one participant observed an increased occurrence of such fraudulent SMS around holidays or birthdays. Participant 11(P11) fell for such fraud and clicked on such message under the assumption that it pertained to an awaited package. P11 stated:

"Yeah, so I fell for this because I actually did have a package coming at the time....it depends on the context and when you are expecting some package...This again happened with three days earlier and I also thought that this was legit at first."(P11)

Also, the combination of curiosity, the anticipation of a package, and a lack of awareness about this type of SMiShing attack can serve as motivations for users to click on such links.

4.2.2 Financial Deception

Of the 25 participants, 36% disclosed instances of financial fraud SMS. Primarily, these messages impersonated banks (5 out of 9), focusing on transaction verifications, debit card lock alerts, and similar themes. P25 provided insights:

"I got a transaction alert message...I looked at the email it was sent from and I was like, that doesn't look right....I was still was nervous about it. So I double checked with my bank. I called them......asked if there was some transaction verification? Because I do bank with Wells Fargo. So I was like I wanted to just double check to make sure."(P25)

Additionally, 2 cases were associated with bill payments, 2 with cryptocurrency offers, and one involving an account block alert. Notably, P16 received MetaMask cryptocurrency wallet alerts, despite not having an account with any crypto wallets.

4.2.3 Fraudulent Business Promotion

Approximately 28% (7 out of 25) received fraudulent business promotion messages. While some promotions were legitimate, participants tended to recall expecting such communications. Vague information, suspicious or unofficial links, and attached images were red flags. P16 shared an example from "K'A'Y" Jewelry, stating:

"Using special characters make it more suspicious....like when they start using characters that are not letters..... I have one in my phone actually that I thought was like, kind of funny. Like pretending to be Kay Jewelers."(P16)

Figure 3 depicts the SMS with three key indicators, as explained by P16 why they believed it was a fraudulent message.

It is worth noting that none of our participants replied to or took any actions in response to unknown business promotions. This is because the distinction between spam and scam is somewhat unclear in their minds; they tend to perceive both as fraudulent activities.

4.2.4 Impersonation Tactics and Deceptive Offers

Around 24% (6 out of 25) received SMS employing impersonation tactics, where the sender pretended to be someone else. Common messages included queries like *"Hi, how are you?"* or *"Can you pick me up at the airport at 6 pm?"* or *"I won't be*
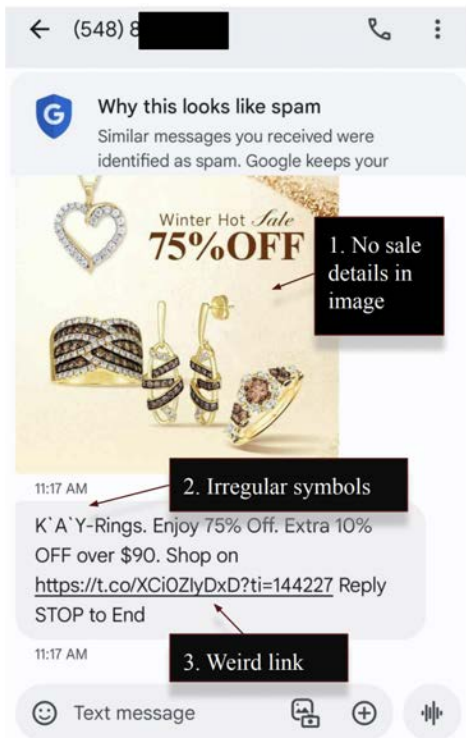
Figure 3: Text message displaying a deceptive business sale offer, highlighting three major suspicious cues identified by Participant 16: absence of detailed sale information in the attached image, irregular symbols used, and a suspicious-looking link

Figure 4: Deceptive SMS impersonating AT&T officials

able to go to the winey with you tomorrow". Some participants chose to ignore such messages, while others, exemplified by P2, initially responded, later recognizing the fraudulent nature of the attempt to gather personal information. P2 shared the tendency to respond stems from the belief that the sender might have dialed a wrong number and is genuinely trying to reach someone.

Another participant received a fake job offer but promptly dismissed it due to the exorbitant amount of money promised in the SMS. Additionally, one participant (P18) fell victim to a gift card fraud while expecting a legitimate gift card. Figure 4 shows the SMS related to this incident. P18 shared their experience: *"I noticed that a group of malicious people. They noticed that I'm going to receive the gift cards, okay?.... In the coming few weeks they called me. And they also send me a text message..... And, to be honest, I first I trusted because I think and they are not asking the age and the gift card numbers or pin numbers. Instead, they're asking whether you have received your gift cards or not"*(P18)

Three participants out of 25 reported receiving fraudulent health and car insurance offers. They noted an increase in health insurance scams after having children, while the car

insurance scams began when they provided their number to auto dealers during a car search.

4.2.5 Political Scams

Regarding political text messages some participants faced confusion distinguishing between legitimate and fraudulent political text messages. Some individuals mistakenly labeled non-harmful messages, specifically related to political campaigns, as fraud or suspicious due to a lack of contextual understanding. Participant P4 exemplified this situation by sharing an SMS, illustrated in Figure 5. P4 shared:

"I think this is a fraud...I don't know who Nikki is, I didn't sign up for that....of course I'm not going to click on the link. Yeah I don't even know the area code "337" and where that number comes from..."(P4)

Moreover, three participants reported receiving political scam SMS. They expressed skepticism about these messages due to irregular lettering and out-of-context content and weird accompanying links.

4.3 Participants' SMS Verification Strategies

We provided three pairs of legit and fraudulent SMS to each participant and requested them to identify them. Our goal was to gain insights into the cues they use when evaluating SMS messages. The comparative chart in Figure 6 shows the participants' capability to differentiate between legitimate and fraudulent SMS within each pair. Specifically, for Pair 2, participants exhibited a tendency to misidentify because they noticed the sentence structure was too informal, and they distrusted SMS messages from shortcodes. Conversely, 3 out of 14 participants incorrectly identified the fraudulent SMS as legitimate. They trusted the 5-digit short code and the instruction to "reply YES to send."
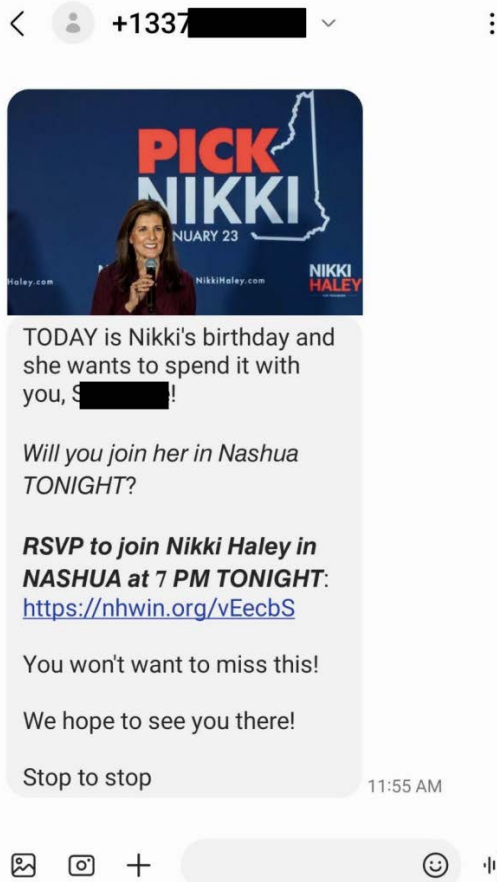
Figure 5: An example of a text message related to a political campaign that raises suspicion due to the absence of contextual understanding and an unfamiliar area code, as reported by P4

Figure 6: Comparative Bar Chart: Participants' Accuracy in Distinguishing Legitimate and Fraudulent SMS. The chart displays participants' performance, with True Positives indicating correctly identified legitimate messages, True Negatives for correctly identified fraudulent messages, False Positives for incorrectly labeled legitimate messages, and False Negatives for misidentifying legitimate messages as fraudulent.

In contrast, their performance significantly improved for Pair 4. The legitimate SMS in this pair followed an official format, including personalized information like the last 4 digits of a card, mention of the mobile app, and reliance on the 5-digit short code. On the other hand, the fraudulent SMS came from an email address, raising suspicion. The format and the link appeared unofficial and dubious to the participants. The following sub-sections discuss the specific cues they consider when assessing text messages.

4.3.1 Cues for Suspicious SMS

Here, we elaborate on the cues for suspicious SMS messages as identified by participants, listing them in descending order from the most mentioned to the least:

Suspicious Contents: 28 out of 29 participants (96.5%) emphasized the significance of text message content in their assessments. Specifically, within the content, 17 participants flagged any SMS containing **links** as suspicious.

P21 said: *"...basically any link telling me to click on this im-*

mediately makes me suspicious..."

Among the SMS pairs presented, pairs 1, 3, 4, 5, and 6 contained links. When both legit and fraud SMS messages included links, participants scrutinized the details of the URLs more closely. They tried to be sure whether the domain matched the official domain of the respective company or service. For pair 1, 13 out of 16 participants correctly identified the fraudulent message. They pointed out the presence of "tinyurl" in the link and the lack of security with "http:" at the beginning as red flags. Pairs 4 and 5 were comparatively easy for our participants. Especially, since the fraudulent link in Pair 4 lacked "https:". However, in Pair 6, many participants found it challenging to identify the legitimate SMS, with 6 out of 13 incorrectly labeling the legitimate message as fraudulent due to the inclusion of random letters and multiple links in a single SMS.

15 out of 29 participants indicated that they primarily examine the content first, assessing for the presence of links, correct grammar, or spelling errors. P24 mentioned:

" I usually glance at it and like the first thing I pick out is spelling errors. If things are misspelled that usually it's like a telltale sign. It's something fishy...I generally just ignore ones that include links."(P24) 8 participants expressed suspicion towards SMS related to money, while 2 individuals noted concern with personal inquiries, and 1 with generalized SMS.

Unofficial Format: 15 out of 29 (51.7%) participants identified an unofficial format, including wrong spelling and grammar (6/15), the use of irregular special characters (5/15), and SMS containing wrong or weird company names, as suspicious. P17 shared:

"I never click on like the suspicious links or misspelled things.....Or like lower-cases yeah, stuff like that....I look into

the format as well, like, how their constructing the sentence if it doesn't read right stuff like that."(P17)

In Pair 2, a significant reason why 7 out of 14 participants incorrectly identified the legitimate SMS as fraudulent was due to unofficial format of that SMS. P17 added:

"The spelling of 'Zelle' starts with a lower-case 'z'..it's definitely fraud...also the '.00' after the money amount is not normal I think"(P17)

Unknown Sender: While judging an SMS, 18 participants mentioned examining the sender's number, email, or short code. However, it is important to note that they do not automatically categorize unknown senders as fraudulent. Instead, they assess the number in various ways. Of these, 3 participants specifically emphasized checking the area code, expressing reluctance to respond if it fell outside their familiar geographic area. P16 said:

"I'll look at the area code but I'll also obviously look at the content....I don't recognize if it's not from somebody in my area code I usually don't answer it."(P16)

11 out of 29 (37.9%) participants considered any SMS from an unknown sender suspicious, especially if it originated from an international number, according to 2 participants. One participant mentioned mistrust towards SMS sent from email and this made them identify the fraud SMS in Pair 4 correctly. On the other hand, there was a level of skepticism towards short codes as senders, resulting in 2 participants mistakenly labeling the legitimate SMS in Pair 1 as fraudulent.

Out of Context SMS: 6 out of 29 participants deemed any out-of-context SMS suspicious. Out-of-context SMS refers to messages that are unexpected or unrelated to the recipient's recent activities or communications, making them appear unusual and potentially fraudulent.

Immediate Action: 4 out of 29 participants stated that any SMS requesting immediate action would be deemed suspicious by them. P21 said:

"... anything has a time frame telling me that I need to respond within one or two hours makes me suspicious.."(P21)

In Pair 6, three participants noted that the mention of the phrase "next 24 hours" led them to believe it was a fraudulent SMS, although it was actually a legitimate one. On the other hand, in Pair 5, the fraudulent SMS contained the phrase "within 4 hours," aiding participants in identifying it as fraudulent.

4.3.2 Cues for Legitimate SMS

Among the cues that contribute to making text messages more reliable and legitimate for our participants, several factors were highlighted. The top factor was making the SMS more specific with personalized information known only to the subscribed business and not to the general public. Other significant cues included a known context, a familiar sender, and an official format. These factors are discussed in detail below:

Contains Personalized Info: The importance of personalized information in SMS was emphasized by 14 participants. It worked as a key indicator in determining the legitimacy of text messages, especially for Pair 3 and 4. In these pairs, 10 out of 14 and 14 out of 17 participants successfully identified the legitimate message, respectively, attributing their success to the presence of personalized details. Participant 5 shared:

"I will trust the second one as it has the last 4 digits of my card... I think it's hard for scammers to get this information"(P5).

Known Context: Next significant indicator was a known context, as highlighted by 11 participants. They expressed trust in SMS messages they were already expecting. Especially, P12 and P13 mentioned deleting any message on their phone if they do not recognize the sender or the reason for texting. P12 further explained that they ask people to call if necessary but not to send text messages.

Known Sender: Ten participants mentioned that they trust SMS messages from known senders. P16 said:

"I usually trust the numbers that are already saved in my phone... or you know, who gives the name and say like, 'Hi, I am Alex, we met at the college today'... like that... so I can relate."

Official Format: Mentioned by 8 participants, having an official format emerged as another key factor. This played a crucial role in the higher success rate in correctly identifying legitimate SMS in Pair 3 and 4. P7, who saw Pair 4, mentioned noticing the $<\#>$ sign in SMS from Bank of America. P7 said:

"The example with Bank of America, there was a the pound sign at the beginning of the message....I feel like I've seen bank messages that also use symbols in the beginning that could be used the key identifier for legitimacy."(P7)

However, the absence of an official format in the legitimate SMS in Pair 2 and 6 posed challenges for participants. For instance, when evaluating SMS Pair 6, P8 mentioned:

"the text is too long..and it is from some peronal phone number..I do not think it is legit, it's fraud"(P8)

Additionally, participants expressed trust in SMS that require no action, involve no personal inquiries, and exhibit correct spelling and grammar.

4.3.3 Visual Indicators

The majority of our participants did not focus on visual indicators when assessing the credibility of an SMS. On the contrary, they identified excessive use of emojis, excessive exclamation marks or dollar symbols, and messages that were either too long or too short as red flags. Android users mentioned that they would be more suspicious of an SMS if it triggered warning signs from the Android SMS spam filters [36]. In contrast, iPhone users, lacking such warning signs, did not anticipate any indicators for fraudulent SMS.

It was noteworthy to observe that in Pair 5 and 6, despite the fraud SMS displaying warning signs according to the

Google Message Spam Filter, iPhone users overlooked the indicators, whereas all Android users were able to identify them. Additionally, in Pair 6, one participant (P18) placed trust in the fraudulent SMS as legitimate due to the presence of a padlock sign for secured RCS chat [44], as provided by Google. This underscores the need for more efficient and effective design in messaging platforms that align with users' mental models [7, 41].

4.4 Verification Behavior

To understand our participants verification behavior we asked them, *"Where do you turn to for verification when you receive a suspicious text message?"* A majority (72.4%) of participants shared that they would contact the bank or the company mentioned in the SMS for confirmation. To do so, they would either use mobile apps or visit the official website to find the correct contact number and verify the authenticity of the received SMS.

Interestingly, two participants shared a unique approach, stating that they would initially consult their father or elder brother for verification, trusting them as the best option. P15 elaborated, stating,

"Well, I go to my dad first and call him to ask about it. If he says it's legit, then I'll call the company. If both sources confirm its legitimacy, then I will trust it."(P15)

Two participants indicated they rely on their own judgment or analytical skills to verify suspicious SMS. Notably, only one participant (P2) mentioned occasionally using ChatGPT to verify links on suspicious SMS.

P2 explained, *"Sometimes I use websites or ChatGPT. For example, around four months ago, I wanted to purchase tickets from an unfamiliar source, and I was unsure about its legitimacy. I asked ChatGPT, 'Is it legal? Do you have any information about the website?' And it told me, 'Yes, it's legit'."*

Two participants mentioned that they do not engage in any verification process. In contrast, only one participant (P27) employs websites such as SpyDialer [48] to verify unknown caller or sender numbers.

4.5 Reporting Behavior

In order to investigate participants' actions upon receiving suspicious or fraudulent SMS messages and their reporting behavior, we asked them the following questions: *"What do you do when you receive a fraudulent text? Have you ever reported it? How? What were your expectations after reporting? And what actually happened?"*

In response, 55.2% (16/29) of our participants indicated that they generally do not report such messages. Among these, five individuals mentioned simply ignoring the message. Four participants actively delete the messages but refrain from reporting. Additionally, five participants rely on the filtering system on their phones and do not check the spam folder to

report such messages. On Android phones, the SMS filter displays a warning sign and provides an explanation for marking a message as spam, as highlighted in green color in Figure 7. However, several participants mentioned that if there was an easier reporting option, as suggested in Figure 7, they would report more.
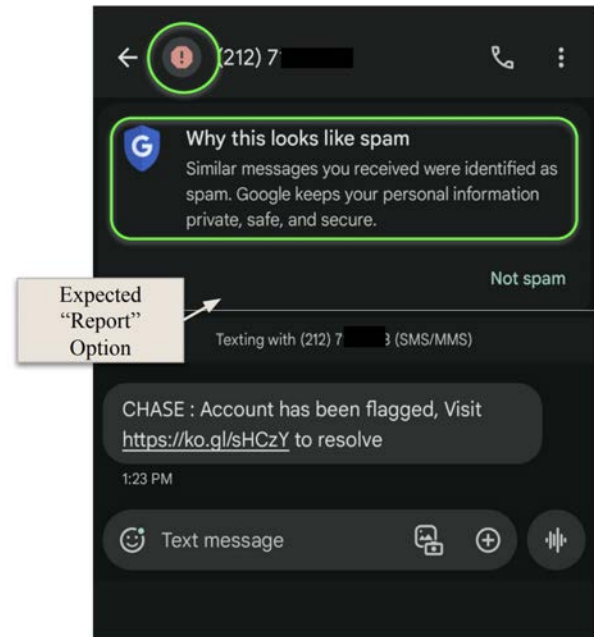
Figure 7: Green-highlighted areas show the warning signs by Android SMS Spam Filters that were appreciated by our participants. Yet, users expressed a desire for more accessible reporting options, as indicated in the image.

Seven out of the 29 participants mentioned that they occasionally report SMS messages, especially if they find them to be alarming, containing sensitive information, or referencing potential financial harm. One participant stated they would only report work-related messages, believing that such reporting could benefit their colleagues.

Six participants shared that they utilize the *"Report Junk"* option on their iPhones when reporting suspicious messages as showed in Figure 8.

iMessage currently lacks spam filters and warning signs, unlike Android. Eight participants suggested that incorporating warning signs, as shown in Figure 8, in suspicious text messages would aid in identification. Interestingly, none of the participants have received any feedback regarding the outcome of their reports. Every participant expressed that they would like to have some feedback or assurance. This feedback would contribute to a sense of confidence and assurance that authorities are actively addressing fraudulent SMS concerns.

Figure 8: The 'Report Junk' option, highlighted in green on the iMessage interface proved useful for reporting by our participants. Also, they expressed a need for warning signs in the indicated area to better identify potential fraud SMS.

4.6 Effect of Prior Cyber Security Training

Among 29 participants 14 individuals(48.3%) reported not having received any formal training or education related to cybersecurity. While not statistically proven, we observed a trend indicating that participants more susceptible to incorrectly identifying fraudulent and legitimate SMS messages belonged to this group without cybersecurity or awareness training, including P8, P15, and P29 (3 out of 29). Interestingly, those who had some form of training demonstrated better abilities in identifying fraudulent messages. Also, these individuals showed heightened suspicion to legitimate messages, often classifying them as potentially fraudulent due to their added caution. Some participants (P10, P22, P26) mentioned attending seminars on cybersecurity during middle or high school.

Two participant mentioned learning through online awareness posts, while another expressed reliance on their analytical skills, saying that formal training was unnecessary to them. Six participants had a major in computer science and were well aware of cybersecurity. Five received occasional security training at their jobs. One participant had received training on cybersecurity at school. Additionally, two participants learned about cybersecurity from their fathers working in the IT industry. They did quite well in identifying the legit and fraud SMS. They expressed that more structured training would enhance their ability to identify patterns in fraudulent SMS.

5 Discussion

5.1 Key Insights

To reduce susceptibility to SMiShing, it is crucial to understand the cues that make SMS messages appear more legitimate or suspicious. Our study explores these cues and investigates the considerations users take into account when making judgments. We found that the presence of URLs or links, unofficial format, immediate action cues are some of the key factors that people find suspicious which aligns with previous SMiShing related studies [15]. Moreover, participants in our study placed significant importance on context when assessing SMS messages. Before evaluating the content or authority of the message, they considered whether they were expecting the text message. Goel et al. noted that exploiting this context as a human weakness is applicable to SMS phishing as well [24]. Furthermore, participants consistently emphasized the significance of personalized information within the content, highlighting its impact on their judgment of the message's legitimacy.

Another significant factor is the participants' frequent confusion between spam and scam text messages [34]. Many businesses use text messages for promotions, but inconsistencies in numbers and formats often lead people to consider these messages as scams. This observation resonates with the idea of implementing improved designs and educational initiatives to facilitate users in distinguishing between legitimate and fraudulent messages [6, 27].

Moreover, participants stressed the importance of enhanced filtering mechanisms on iPhones to identify and block fraudulent SMS. This recommendation aligns with the growing need for adaptive and advanced security features in mobile devices to proactively detect and prevent SMiShing attacks. Our study also shows that participants often have difficulties reporting incidents because they are not familiar with the "7726" reporting service [3] and feel the need for a more accessible reporting option. This emphasizes the importance of making reporting processes simpler to ensure fast and efficient reporting of fraudulent SMS.

While our study did not identify specific patterns related to mobile carrier, background, or job influencing susceptibility to SMS phishing, we found that younger participants aged 18 to 24, especially those without prior cybersecurity training, were more vulnerable. This aligns with a recent survey by Faklaris et al. on US demographics [22]. Interestingly, some younger participants in our study excelled in identifying fraud vs. legit SMS, mentioning prior cybersecurity training. Additionally, older individuals exhibited greater caution when receiving SMS, often due to security training at work. These findings underscore the potential impact of educational initiatives in enhancing users' ability to identify and mitigate SMS phishing threats.

Our study not only contributes valuable insights into the

nuances of SMS phishing but also advocates for addressing aspects such as improved filtering mechanisms, accessible reporting options, and comprehensive training programs. These are essential to fortifying users against evolving SMiShing threats in the digital landscape.

5.2 Finding Alignment: Comparing SMiShing and Email Phishing

Our study reveals significant parallels between SMiShing and email phishing, particularly in how users assess the legitimacy of messages. Participants placed considerable importance on context when evaluating SMS messages, aligning with findings from prior research on email phishing [8, 24, 30].

Moreover, our study participants consistently highlighted the significance of personalized information within the content of SMS messages, a factor similarly emphasized in email phishing [30, 41]. However, unlike email phishing, where users can hover over links to verify their legitimacy , this action is not feasible on mobile devices. This difference underscores the sophistication of SMiShing attacks, as perpetrators tailor messages to appear genuine by incorporating recipient-specific details and an official-looking format.

Participants also indicated that security symbols, such as padlock icons or indications of secured SMS, are trust indicators. This finding aligns with Jakobsson (2007), who noted the importance of such symbols in establishing perceived safety in email communication [30]. Similarly, users in our study found it helpful to have warning signs on Android phones for potential fraud in SMS, just like those used in email systems. This supports findings from research on email phishing [41].

However, there are notable differences between SMiShing and email phishing. For instance, SMS messages typically do not feature logos or third-party endorsements, which are recognizable brand logos or verifications from credible third parties, commonly used to signify legitimacy in email phishing [30]. This absence of visual and endorsement cues in SMS requires users to rely more heavily on other indicators such as context, content personalization, and security symbols.

These insights indicate that while there are significant overlaps in how users perceive and respond to SMiShing and email phishing, the unique constraints and characteristics of SMS messaging necessitate different strategies for identifying and mitigating these threats. Our findings emphasize the need for tailored approaches to enhance user awareness and defenses against SMiShing.

6 Limitations

We conducted interviews with 29 participants in a major southeastern U.S. city, providing valuable insights into the cues they consider when evaluating the legitimacy of text messages and enhancing our understanding of their experiences with SMiShing. The participants, diverse in age and profession, may not fully represent other locations, particularly rural or underdeveloped areas where awareness levels differ, potentially introducing bias. We used a variety of SMS visual styles: Pairs 1-4 mimicked the iMessage UI, while Pairs 5-6 presented real Android examples. This diversity revealed that Android users recognized warnings more effectively due to familiarity with red indicators, though the visual dissimilarity between pairs is a noted limitation. The selection process was pseudo-random and counterbalanced to ensure exposure to different pairs, but not all participants saw each visual style. Additionally, focusing on financial SMiShing examples may not encompass the full range of SMiShing attacks, suggesting future research should include a broader spectrum. The recruitment text aimed to engage individuals familiar with or open to discussing fraud messages, which may have introduced bias. Future studies should consider a wider variety of SMiShing categories and acknowledge the possibility of emerging SMiShing patterns.

7 Conclusion

Our study sheds light on the pressing issue of SMS phishing (SMiShing) and its significant impact on individuals. Through exploring real-life experiences, we gained nuanced insights into susceptibility factors, with participants highlighting cues crucial for distinguishing between legitimate and fraudulent SMS - emphasizing personalized information, known senders, and official formats. Furthermore, our study contributes valuable recommendations for telecom and mobile companies to enhance security measures. By proposing design suggestions informed by user feedback, we aim to empower these entities to better protect their users from falling victim to SMS phishing scams. As the first qualitative exploration into SMiShing, our research advances the understanding of this cybersecurity challenge. The findings underscore the need for proactive measures and heightened awareness to mitigate the risks associated with fraudulent SMS. In an era where digital communication plays a pivotal role, safeguarding users against SMS phishing is imperative for fostering a secure and trustworthy mobile communication environment.

Acknowledgments

References

[1] 2023 data breach investigations report. Retrieved: February 11, 2024 from https://www.verizon.com/business/resources/reports/dbir/.

[2] Recognizing bank account fraud: Bank of america. Retrieved: February 10, 2024 from https://www.bankofamerica.com/security-center/faq/sharing-information/.

[3] How to recognize and report spam text messages, Jun 2022. Retrieved: February 10, 2024 from https://consumer.ftc.gov/articles/how-recognize-and-report-spam-text-messages.

[4] Threat report 2023: State of the phish, Aug 2023. Retrieved: February 10, 2024 from https://www.proofpoint.com/us/resources/threat-reports/state-of-phish.

[5] Consumer sentinel network data book 2023, Feb 2024. Retrieved: February 11, 2024 from https://www.ftc.gov/reports/consumer-sentinel-network-data-book-2023.

[6] Elham Al Qahtani, Yousra Javed, Sarah Tabassum, Lipsarani Sahoo, and Mohamed Shehab. Managing access to confidential documents: A case study of an email security tool. *Future Internet*, 15(11):356, Oct 2023.

[7] Auwal Shehu Ali and Zarul Fitri Zaaba. Mental models review for security and privacy policy: An approach. In *2021 International Conference on Information Technology (ICIT)*, pages 905–909. IEEE, 2021.

[8] Zainab Alkhalil, Chaminda Hewage, Liqaa Nawaf, and Imtiaz Khan. Phishing attacks: A recent comprehensive study and a new anatomy. *Frontiers in Computer Science*, 3:563060, 2021.

[9] Mohamed Alsharnouby, Furkan Alaca, and Sonia Chiasson. Why phishing still works: User strategies for combating phishing attacks. *International Journal of Human-Computer Studies*, 82:69–82, 2015.

[10] APWG. Phishing activity trends report, 2nd quarter 2023, Nov 2023.

[11] Mark Blythe, Helen Petrie, and John A Clark. F for fake: four studies on how we fall for phish. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 3469–3478, 2011.

[12] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

[13] Chuck Brooks. Cybersecurity trends statistics for 2023; what you need to know, Sep 2023. Retrieved: February 10, 2024 from https://www.forbes.com.

[14] Emily Cahill. Phishing, smishing and vishing: What's the difference?, Jul 2023. Retrieved: February 12, 2024 from https://www.experian.com/blogs/ask-experian/phishing-smishing-vishing/.

[15] Max Clasen, Fudong Li, and David Williams. Friend or foe: An investigation into recipient identification of sms-based phishing. In *Human Aspects of Information Security and Assurance: 15th IFIP WG 11.12 International Symposium, HAISA 2021, Virtual Event, July 7–9, 2021, Proceedings 15*, pages 148–163. Springer, 2021.

[16] Shelby R Curtis, Prashanth Rajivan, Daniel N Jones, and Cleotilde Gonzalez. Phishing attempts among the dark triad: Patterns of attack and vulnerability. *Computers in Human Behavior*, 87:174–182, 2018.

[17] Rachna Dhamija, J Doug Tygar, and Marti Hearst. Why phishing works. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 581–590, 2006.

[18] Xun Dong, John A Clark, and Jeremy L Jacob. User behaviour based phishing websites detection. In *2008 International Multiconference on Computer Science and Information Technology*, pages 783–790. IEEE, 2008.

[19] Julie S Downs, Mandy Holbrook, and Lorrie Faith Cranor. Behavioral response to phishing risk. In *Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit*, pages 37–44, 2007.

[20] Julie S Downs, Mandy B Holbrook, and Lorrie Faith Cranor. Decision strategies and susceptibility to phishing. In *Proceedings of the second symposium on Usable privacy and security*, pages 79–90, 2006.

[21] Serge Egelman, Lorrie Faith Cranor, and Jason Hong. You've been warned: an empirical study of the effectiveness of web browser phishing warnings. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1065–1074, 2008.

[22] Cori Faklaris, Heather Richter Lipford, and Sarah Tabassum. Preliminary results from a us demographic analysis of smish susceptibility. *arXiv preprint arXiv:2309.06322*, 2023.

[23] Paul Gillin. The history of phishing, Jan 2021. Retrieved: February 11, 2024 from https://www.verizon.com/business/resources/articles/s/the-history-of-phishing.

[24] Sanjay Goel, Kevin Williams, and Ersin Dincelli. Got phished? internet security and human vulnerability. *Journal of the Association for Information Systems*, 18(1):2, 2017.

[25] Yasmeen Hanif and Harjinder Singh Lallie. Security factors on the intention to use mobile banking applications in the uk older generation (55+). a mixed-method study using modified utaut and mtam-with perceived cyber security, risk, and trust. *Technology in Society*, 67:101693, 2021.

[26] Dermot Harnett and W. Stuart Jones. Smishing vs. phishing: Understanding the differences: Proofpoint us, May 2023. Retrieved: February 12, 2024 from https://www.proofpoint.com/us/blog/email-and-cloud-threats/smishing-vs-phishing-understanding-differences.

[27] Cormac Herley. So long, and no thanks for the externalities: the rational rejection of security advice by users. In *Proceedings of the 2009 workshop on New security paradigms workshop*, pages 133–144, 2009.

[28] Jason Hong. The state of phishing attacks. *Communications of the ACM*, 55(1):74–81, 2012.

[29] Markus Jakobsson. Two-factor inauthentication–the rise in sms phishing attacks. *Computer Fraud & Security*, 2018(6):6–8, 2018.

[30] Markus Jakobsson, Alex Tsow, Ankur Shah, Eli Blevis, and Youn-Kyung Lim. What instills trust? a qualitative study of phishing. In *Financial Cryptography and Data Security: 11th International Conference, FC 2007, and 1st International Workshop on Usable Security, USEC 2007, Scarborough, Trinidad and Tobago, February 12-16, 2007. Revised Selected Papers 11*, pages 356–361. Springer, 2007.

[31] Mohammad S Jalali, Maike Bruckes, Daniel Westmattelmann, and Gerhard Schewe. Why employees (still) click on phishing links: investigation in hospitals. *Journal of medical Internet research*, 22(1):e16775, 2020.

[32] Simon Kemp. Digital 2023: Global overview report - datareportal – global digital insights, Feb 2023. Retrieved: February 02, 2024 from https://datareportal.com/reports/digital-2023-global-overview-report.

[33] Shahedul Huq Khandkar. Open coding. *University of Calgary*, 23(2009):2009, 2009.

[34] Alex Kigerl. Spam-based scams. *The Palgrave Handbook of International Cybercrime and Cyberdeviance*, pages 877–897, 2020.

[35] Stewart Kowalski and Mikael Goldstein. Consumers' awareness of, attitudes towards and adoption of mobile phone security. In *20th International Symposium on Human Factors in Telecommunication*, pages 20–23, 2006.

[36] Neil Kumaran. New gmail protections for a safer, less spammy inbox, October 2023. Retrieved: February 12, 2024 from https://blog.google/products/gmail/gmail-security-authentication-spam-protection/.

[37] Keepnet Labs. Smishing statistics 2023: The latest trends and numbers in sms phishing, Jan 2024. Retrieved: February 10, 2024 from https://keepnetlabs.com/blog/smishing-statistics-2023-the-latest-trends-and-numbers-in-sms-phishing.

[38] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–23, 2019.

[39] Tanya McGill and Nik Thompson. Old risks, new challenges: exploring differences in security between home computer and mobile device use. *Behaviour & Information Technology*, 36(11):1111–1124, 2017.

[40] Aleksandr Nahapetyan, Sathvik Prasad, Kevin Childs, Adam Oest, Yeganeh Ladwig, Alexandros Kapravelos, and Bradley Reaves. On sms phishing tactics and infrastructure. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 169–169. IEEE Computer Society, 2024.

[41] Justin Petelka, Yixin Zou, and Florian Schaub. Put your warning where your link is: Improving and evaluating email phishing warnings. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–15, 2019.

[42] Md Lutfor Rahman, Daniel Timko, Hamid Wali, and Ajaya Neupane. Users really do respond to smishing. In *Proceedings of the Thirteenth ACM Conference on Data and Application Security and Privacy*, pages 49–60, 2023.

[43] IBM Research. What is smishing (sms phishing)?, 2023. Retrieved: February 12, 2024 from https://www.ibm.com/topics/smishing.

[44] Drew Rowny. Access the assistant in messages, plus the latest on rcs, Feb. 2019. Retrieved: February 09, 2024 from https://blog.google/products/messages/access-assistant-messages-plus-latest-rcs/.

[45] Johnny Saldaña. The coding manual for qualitative researchers. *The coding manual for qualitative researchers*, pages 1–440, 2021.

[46] Steve Sheng, Mandy Holbrook, Ponnurangam Kumaraguru, Lorrie Faith Cranor, and Julie Downs. Who falls for phish? a demographic analysis of phishing susceptibility and effectiveness of interventions. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 373–382, 2010.

[47] Steve Sheng, Bryant Magnien, Ponnurangam Kumaraguru, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. Anti-phishing phil: the design and evaluation of a game that teaches people not to fall for phish. In *Proceedings of the 3rd symposium on Usable privacy and security*, pages 88–99, 2007.

[48] Phillip Tracy. How to do a reverse phone number lookup without paying a dime, March 2021. Retrieved: February 11, 2024 from https://www.dailydot.com/debug/reverse-phone-lookup/.

[49] Melanie Volkamer, Karen Renaud, Benjamin Reinheimer, and Alexandra Kunz. User experiences of torpedo: Tooltip-powered phishing email detection. *Computers & Security*, 71:100–113, 2017.

[50] Heather Young, Tony van Vliet, Josine van de Ven, Steven Jol, and Carlijn Broekman. Understanding human factors in cyber security as a dynamic system. In *Advances in Human Factors in Cybersecurity: Proceedings of the AHFE 2017 International Conference on Human Factors in Cybersecurity, July 17- 21, 2017, The Westin Bonaventure Hotel, Los Angeles, California, USA 8*, pages 244–254. Springer, 2018.

A   Appendix

A.1   Recruitment Script

This section describes the text we used for our recruitment process. This text was adapted for use in email campaigns, flyers, websites, and social media advertisements to recruit participants for the study:

Receive $25 if Selected for Our Research Study on SMiShing!

We are conducting a research study on SMiShing (fraudulent text messages) and are seeking participants to help us gain deeper insights into this issue.

About the Study: Our study aims to understand how individuals perceive and respond to suspicious or fraudulent text messages. SMiShing, or SMS phishing, is a growing cybersecurity threat, and your experiences and opinions can greatly inform our research.

Participant Criteria: We are looking for diverse participants who meet the following criteria:

- Aged 18 or older residing in Charlotte Metro area

- Use a mobile phone

- Can attend an in-person interview and bring their mobile phone

Study Details: The study will be an in-person interview at/near the UNCC campus. We kindly request that you bring your mobile phone to the interview session. You will be asked if you are willing to share any suspicious or fraudulent text messages that you have received on your phone. The interview is expected to last no more than one hour, and upon completion, you will be rewarded with a $25 Amazon e-gift card.

By participating in this study, you will help us develop a deeper understanding of SMiShing, its impact, and how individuals can protect themselves from such threats. Your insights will contribute to improved cybersecurity practices and awareness.

How to Get Involved: If you are interested in participating, please complete a brief eligibility survey to determine your eligibility and provide your contact information.

If you have any questions about the study, please contact Sarah Tabassum or Faculty Advisor, Dr. Cori Faklaris, at stabass2@charlotte.edu or cfaklari@charlotte.edu.

Thank you for your time and help!

A.2   Interview Guide

Greetings and Introduction: At the start of the study, participants will be greeted and introduced to the research topic. [Good Morning/afternoon. Thank you so much for participating in our study. The research topic is centered around identifying and understanding fraudulent text messages. We are doing this study to understand how users can identify fraud text vs. real text. We want to know about your experience and your opinion about such text messages. We will record the audio responses of yours. And for some questions, I will ask you to think aloud. Is that okay with you? Okay, let's get started then.]

Interview: During the interview, the following questions will be asked:

1. How frequently do you use your mobile phone for texting? What about the texting app on your phone?

2. Can you let me know what type of phone you use? For example, Apple, Android?

3. Which company provides your mobile service? [For example, are you with AT&T, Verizon, or Mint?]

4. Have you faced any privacy or security concerns related to your phone?

5. Have you ever received any fraudulent or suspicious text messages? Especially in the last three months? If yes, What was that? [At this point we will ask them if they can show us any such SMS on their phone. If they can, we will ask them to take a screenshot of that suspicious/fraud SMS and share with us.]

6. What did you do about it? Why? What influenced your decision?

7. Now, I will show you some Real and Fraud text messages to you for understanding your perception and decision-making process. [We want to understand the decision-making Process for each text message] For each text message we will tell them, "Describe the process you go through when deciding whether to trust a text message or not".

8. When you receive a text message from an unfamiliar source, what are the first things you notice or look for?

9. Is there anything specific in a text message that makes you suspicious? For example, Are there any specific words, phrases, and visual elements that trigger suspicion?

10. Is there anything specific in a text message that makes it appear legitimate? For example, Are there any words, phrases, and visual elements that enhance legitimacy?

11. Do you pay attention to symbols or indicators (like a green checkmark, blue icon, or yellow warning sign) when judging the credibility of a text message? If so, please explain how these indicators influence your trust or suspicion

12. Have you received an SMS with such visual cues, and if so, how did they impact your perception of its legitimacy?

13. Are there specific colors, symbols, or icons in an SMS that make you more or less suspicious?

14. Where do you turn to for verification when you receive a suspicious text message? [Specific reasons for your choice and any phone features used for verification will be discussed.]

15. Have you received formal training or education on computer security/ cyber security or text message security? If so, please describe its effectiveness in preparing you to detect and respond to text message fraud, attacks, or spam

16. What do you do when you receive a fraudulent text? Why?

17. Do you report it? How do you report it? Why?

18. What are your expectations after reporting? And what actually happened?
Now we will ask some questions for feedback:

19. How do you think text message interfaces could be improved to help users identify fraudulent text messages better?

20. Is there anything else you'd like to share on this topic?

[Thank you so much. That's all I had to share today. We really value your thoughts and involvement. I will stop the recording now.] Additionally, participants will be provided with a document on "Best Practices to Identify Fraudulent Text Messages" for their reference.

A.2.1 Best Practices to Identify Fraudulent Text Messages

- Be Skeptical of Unknown Senders: Avoid clicking on links or responding to messages from unknown or suspicious senders.

- Double-Check the Sender's Information: Verify the sender's identity by cross-referencing contact details with official sources.

- Beware of Urgent Requests: Be cautious if the message conveys a sense of urgency, asking for immediate action or personal information.

- Verify Web Addresses (URLs): Scrutinize any links provided in text messages. Confirm the legitimacy of the website before clicking.

- Look for Spelling and Grammar Errors: Fraudulent messages may contain typos, incorrect grammar, or unusual language.

- Avoid Sharing Personal Information: Never share sensitive personal or financial information through text messages.

- Stay Informed about Scams: Keep up-to-date with common text message scams and fraud tactics to recognize red flags.

- Use Official Contact Channels: If you receive a message from a bank, government agency, or service provider, contact them through official channels to verify its authenticity.

- Enable Two-Factor Authentication (2FA): Use 2FA whenever possible to add an extra layer of security to your accounts.

- Report Suspicious Messages: If you receive a fraudulent text message, report it to your mobile carrier and the appropriate authorities.

- Educate and Share Information: Share knowledge about text message scams with family and friends to collectively protect against fraud.

- Verify Prize Winnings: Be cautious of messages claiming you've won a prize or lottery, as these are often scams.

- Trust Your Instincts

- Regularly Update Your Mobile Device

Remember that text message scams can take various forms, so it's essential to stay vigilant and employ these best practices to protect yourself and your personal information.

A.3 Code Book

Main Code	Definition	Sub-codes	Frequency	Sub-sub-codes	Frequency	Examples from Transcripts
Cues: Suspicious	Any anomaly or irregularity within the message that triggers doubt or concern about its legitimacy, indicating potential fraudulent or malicious intent	Suspicious Content	28	Links	17	"The one thing is the content of messages.. for example, if there is a link, I prefer to check the link. Not clicking, just reading the link, or something like that." [P2]
				Money Related SMS	8	"...if it's related to money, that would definitely become more suspicious." [P9]
				Generalized SMS	1	"..when it looks like a very generalized message, like it's being sent to thousands of people, and it's not personalized to me…" [P3]
				Personal Inquiry	2	"When they ask for my personal information, I become more suspicious." [P11]
		Unofficial Format	15	Wrong/Weird Names	4	"I saw there are some weird Google forms and they have some weird company names." [P8]
				Irregular/Special Characters	5	"...obviously like when they start using characters that are not letters. Okay, you know, and decimals that I have one in my phone.." [P16]
				Grammar/Spelling Error	6	" …some of them sometimes are misspelled and have random capitalization that I noticed." [P17]
		Unknown Sender	11	Any Unknown Sender	7	"Generally, if I don't recognize the sender and the message doesn't align with my daily life or the businesses I usually interact with, that makes me a little suspicious." [P5]
				Email	1	"If I receive a text from an email address, that's usually suspicious. I don't expect banks to use emails for text messages." [P25]
				Intl. Number	2	"I become suspicious if it is coming from some crazy international numbers..." [P27]
				5 digit short-code	1	"There are no strict rules or regulations for these numbers, so it's very easy to mask this type of five-digit code.." [P14]
		Out of Context SMS	6	-	-	"Like, I always ask myself, Was I expecting this SMS? Does it relate to my daily life? If not, I become suspicious." [P7]
		Immediate Action	4	-	-	"Yes, like 'call us now.' Okay? When they say immediate action is required, like things that need to be fixed right away.." [P25]

Main Code	Definition	Sub-codes	Frequency	Sub-sub-codes	Frequency	Examples from Transcripts
Cues: Legit	Trustworthy indicators within a text message that instill confidence in the message's authenticity	Contains Personalized Info	14	-	-	"I guess if it's more personal to me, I'm able to recognize it. Like if they mention something we've discussed before, or if they phrase it as if they've had a conversation with me" [P22]
		Known Context	11	-	-	"If it's from a bank or a place I go frequently, I will trust them." [P17]
		Known Sender	10	-	-	"If I've received texts from them before or if I know them, it's easier to trust. Many companies use the same number for messages. For instance, if it's from Amazon to verify my account, I'll have all our previous texts in that conversation. So if it's a company or bank that I've used before, I'd expect their message to come from the same number." [P12]
		Official Format	8	No Personal Info Inquiry	2	"Like, legitimate senders will never ask for your personal information. If I ordered something, they should already know about my info…" [P8]
				No Action Required	1	"Legitimate SMS will not ask for urgent actions like 'your card is blocked, call now,' or 'go to this link now', you know…." [P10]
				Correct Spelling & Grammar	1	"…if it's coming from a legitimate company or bank, I expect them to have proper grammar and no spelling mistakes, which are usually seen in fraudulent SMS" [P28]
				Correct Format	4	"..usually the companies follow some formats. I look for the correct format while judging legit SMS." [P6]

Main Code	Definition	Sub-codes	Frequency	Sub-sub-codes	Frequency	Examples from Transcripts
Initial Hook	The first thing they notice while judging legitimacy of an SMS from unknown source	Caller ID Information	18	Unknown Number	15	"…I always check the sender's number in the first place or where the SMS is coming from…" [P5]
				Area Code	3	"The area code where it's coming from is the first thing I check…It gives me a quick idea if the message is trustworthy or not." [P16]
		Content	15	Format of SMS	9	"Usually, the format of the text first grabs my attention, whether it's official or vague, you know…" [P27]
				Links	3	"…if it has a link, it gets my attention first…I become more suspicious about the text message…" [P9]
				Grammar/Spelling Error	3	"….always the spelling and grammar... the scammers usually have lots of spelling mistakes, I noticed." [P3]
		Context	5	-	-	"Like, I always ask myself, Was I expecting this SMS? Does it relate to my daily life? If not, I become suspicious." [P7]

"I would not install an app with this label": Privacy Label Impact on Risk Perception and Willingness to Install iOS Apps

David G. Balash
University of Richmond

Mir Masood Ali
University of Illinois Chicago

Chris Kanich
University of Illinois Chicago

Adam J. Aviv
The George Washington University

Abstract

Starting December 2020, all new and updated iOS apps must display app-based privacy labels. As the first large-scale implementation of privacy nutrition labels in a real-world setting, we aim to understand how these labels affect perceptions of app behavior. Replicating the methodology of Emani-Naeini et al. [IEEE S&P '21] in the space of IoT privacy nutrition labels, we conducted an online study in January 2023 on *Prolific* with $n = 1,505$ participants to investigate the impact of privacy labels on users' risk perception and willingness to install apps. We found that many privacy label attributes raise participants' risk perception and lower their willingness to install an app. For example, when the app privacy label indicates that *financial info* will be collected and linked to their identities, participants were 15 times more likely to report increased privacy and security risks associated with the app. Likewise, when a label shows that *sensitive info* will be collected and used for cross-app/website tracking, participants were 304 times more likely to report a decrease in their willingness to install. However, participants had difficulty understanding privacy label jargon such as "diagnostics", "identifiers", "track" and "linked". We provide recommendations for enhancing privacy label transparency, the importance of label clarity and accuracy, and how labels can impact consumer choice when suitable alternative apps are available.

1 Introduction

Smartphone applications (apps) have become a necessary part of most people's daily lives [17, 46, 47], and app marketplaces such as the Apple App Store [5] provide smartphone users the ability to quickly install a plethora of apps to meet their needs. Today's smartphones come with an impressive array of sensors, such as microphones, cameras, GPS, gyroscopes, and accelerometers. These sensors allow apps to collect more types and larger amounts of data from users of smartphones [44], increasing the privacy risks within the mobile environment [2]. Research has shown that smartphone

users are concerned about their privacy when it comes to their mobile apps [4, 29, 39, 56], but are often unaware of the extent of app data collection [25, 37, 40, 48].

To help people overcome the burdens associated with reading privacy policies [20, 33, 49, 54], researchers designed privacy nutrition labels [9, 19, 22, 23, 35, 36, 38, 55, 57] to improve privacy communication and do away with natural language presentations of privacy behavior. Apple privacy labels were introduced in December 2020 [6, 13] to provide users with more transparency about the data being collected by apps [32]. The labels present users with a standardized set of information about the data being collected, such as the type of data (e.g., location, search history), the purpose of the data collection (e.g., targeted advertising, app functionality), and whether the data is linked to the user's identity [1]. These labels aim to help users make more informed decisions about which apps to use and increase trust in the app ecosystem. Labels have the potential to help users make informed choices when selecting an application to install. Therefore, it is important to understand whether privacy labels lead to better privacy outcomes for users such that users' privacy expectations align with the actual behavior of the apps they use.

Our study replicates the methodology and extends the results from Emami-Naeini et al. [23] on IoT device privacy labels into the ecosystem of Apple's iOS privacy labels, a real-world, large-scale (over a million apps) deployment of privacy labels. The methodology across both studies emphasizes comparing consumer reactions to hypothetical products/apps with differing designs and intuitive privacy implications, one with an expectation of higher privacy invasion and one with a lower expectation. Emani-Naeini et al. considered a hypothetical smart lightbulb (lower expectations) and a smart speaker (higher expectations); we compare a hypothetical note-taking app (lower) to a social media app (higher). By extending the prior IoT study to the iOS privacy label ecosystem, we provide both a point of comparison between the two settings and also how privacy labels in iOS, in particular, have the potential to affect consumer behavior.

We conducted an online survey on *Prolific* [52] in January

2023 with $n = 1,505$ participants to measure the effectiveness of privacy labels in conveying privacy risk to users, and the impact labels have on users' willingness to install an application. The survey structure was based on the methodology of Emami-Naeini et al. [23], which looked at how the proposed design for Internet of Things (IoT) labels would influence consumers' purchase decisions of IoT devices. We asked users about their experiences with the privacy labels on the App Store and how these labels impacted their app installation decision-making. These methods allowed us to answer the following research questions:

RQ1 *[App Concern] What experiences and concerns do users have with the apps they have already installed or considered installing?*

When considering social media and note-taking apps, a greater percentage of participants, 62% versus 34%, reported being at least *Somewhat concerned* with how social media apps would use, collect, and store information. Yet more participants reported they had previously installed a social media app than a note-taking app, 88% to 49%. Privacy concerns were more often cited as a reason for not installing a social media app than a note-taking app.

RQ2 *[Understanding of Privacy Labels] How do users understand the data collection information summarized on the privacy label?*

Participants generally understand the meaning of many privacy label data categories and privacy types. However, we found participants had trouble understanding some of the data categories such as *Other data*, *Diagnostics*, and *Identifiers*. There were also issues of understanding with particular jargon such as "track" and "linked", as well as confusion with the terminology such as *Contacts* versus *Contact info*.

RQ3 *[Risk and Willingness to Install] Which app privacy label attributes significantly influence user risk perception and willingness to install and in what ways?*

The *Data not collected* privacy type was the only label attribute that consistently decreased risk perception and increased willingness to install. Most attributes increased risk perception and reduced willingness to install by at least some amount. The attributes that caused the most significant increase in risk perception and decrease in willingness to install were the *Financial info*, *Sensitive info*, and *Browsing history* data categories, and the *Data used to track you* privacy type.

Participants in the study expressed dissatisfaction with the clarity of privacy labels, emphasizing that these labels often needed more detailed information about an app's data collection behavior, making it challenging to gauge its security and privacy risks. This phenomenon is termed the "transparency paradox," [51] where trying to summarize information handling practices, like through privacy labels, might necessarily omit critical details, leading to confusion and mistrust. Striking the right balance between offering summarized information and exhaustive detail is vital for informed user decisions

about privacy. The study also indicated that while privacy labels can reassure users about upfront data collection, they might foster a false sense of security, leading to complacency. There is a need for more effective oversight of privacy label accuracy and consumer education on their limitations.

Furthermore, the availability of alternative apps in the marketplace can influence users' willingness to compromise on privacy. Participants in our study associated data collection with the app's purpose, where incongruences led to reduced trust in the app. Overall, while privacy and security labels have the potential to be influential in shaping user perceptions and decisions, their efficacy relies on their accuracy and completeness, necessitating further research to optimize their design and implementation for a transparent app environment.

2 Background and Related Work

Labels have been used as an effective means to communicate information to end users on products like food (Nutrition Facts) [28] and home appliances [3, 18]. Drawing inspiration from these labels, Kelly et al. [35, 36] developed a privacy label that presents how websites collect, use, and share consumers' personal information. This was later extended [38] in the design of a "Privacy Facts" label for mobile apps. The label detailed information that apps collect along with their intended use. Subsequently, Emami-Naeini et al. [22] developed and evaluated similar labels for Internet of Things (IoT) devices. Over the years, multiple researchers have studied and provided recommendations on designing similar privacy notices from a variety of perspectives [9, 19, 22, 23, 35, 36, 38, 55, 57].

To determine which privacy and security label attributes most impact consumers' risk perception and willingness to purchase Internet of Things (IoT) devices, Emami-Naeini et al. [23] designed a study to measure the effectiveness of each privacy and security attribute-value pair in isolation. This allowed the researchers to assess each attribute's impact and identify misconceptions associated with individual attributes. The study found that attribute values intended to communicate increased risk were generally perceived that way by participants. Still, the study also found risk perception did not always align with willingness to purchase the device. Furthermore, they make recommendations for improving the privacy labels, including reducing information uncertainty (purpose, harms, controls), improving information placement between primary and secondary layers, and reducing misconceptions by providing explanations to consumers.

Other work on Apple's privacy labels. Li et al. [45] interviewed 12 developers and reported their difficulty understanding labels. Gardner et al. [30] developed a tool that analyzes code and prompts developers to report data collection practices in their labels. Kollnig et al. [41] evaluated 1,759 apps before and after the introduction of Apple's App Tracking Transparency and privacy labels. They found instances of
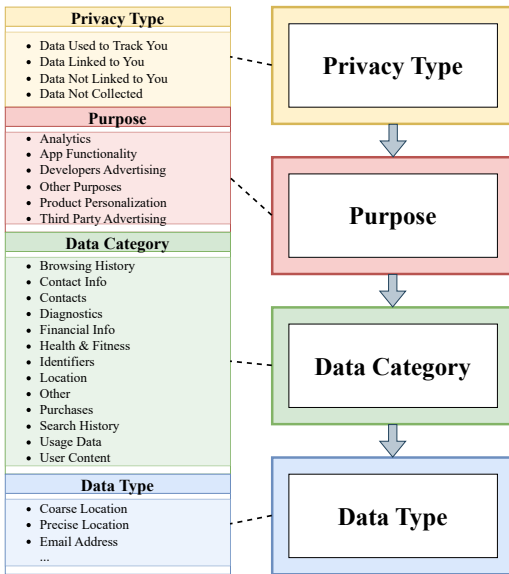
Figure 1: Hierarchical Structure of a Privacy Label.

apps violating Apple's policies and tracking users. In a longitudinal study of privacy label adoption, Balash et al. [8] analyzed weekly snapshots of the App Store for over a year and identified an increase in apps with labels and likely underreporting by developers forced to provide a label on a version update. Xiao et al. [61] analyzed data flows within 5,102 apps and found inconsistencies between app practices and reported privacy labels. Garg and Telang [31] reported a reduction in app demand following privacy label disclosures of the collection of sensitive information. To test the usability of iOS app-based privacy nutrition labels, Zhang et al. [62] conducted an interview study with lay iPhone users. They found dissatisfaction and misunderstandings that reduced the effectiveness of the label, such as confusing structure, unfamiliar terms, and lack of control over permissions settings.

Structure of Apple's Privacy Labels. Apple's privacy labels are similar in structure and content to prior work on privacy nutrition labels [38]. The label follows a hierarchical model (see Figure 1) and describes data collection practices under four levels: **(1) Privacy Type:** Describes how the collected data is handled, i.e., (a) if the data is anonymized, (b) if the data can be used to identify users, and (c) if the data is used to track users (with third parties). An app's privacy label may contain a combination of one, two, or all three of these types. A fourth, mutually exclusive privacy type indicates that the app collects no user data. **(2) Purpose:** Describes the reason for data collection, e.g., for advertising, analytics, etc. **(3) Data Category:** Presents a high-level category for collected data, e.g., *Location*, *Contact Info*, etc. **(4) Data Type:** Granular information under the Data Category, e.g., data types under *Location* can be *Precise Location* and *Coarse Location*.

3 Method

Study Procedure As previously noted, the methods of this study are replicating the work of Emami-Naeini et al. [23] from the IoT privacy nutrition label to Apple's iOS App labels. In Emami-Naeini et al.'s design for IoT, they considered a single label applied to two hypothetical devices: a smart light bulb and a smart speaker. They hypothesized that the light bulb would have low privacy implications with consumers while the smart speaker would have higher privacy implications. This helped them compare participants' associated privacy risk and willingness to purchase in different settings with different privacy expectations.

We replicated their design in the context of iOS apps. Participants viewed two hypothetical apps with different privacy expectations: a note-taking app (less privacy-invasive) and a social media app (more privacy-invasive). Like Emami-Naeini et al., we compared how each privacy label, when individually applied to different settings, affects consumers' willingness to install an app and their associated privacy risks.

Following the design of [23], we considered the hypothetical app as a between-subject factor and the privacy information displayed on the iOS privacy label as a within-subject factor. We randomly assigned each participant to answer questions about 3 of the 43 possible privacy label attributes. Forty-two privacy label attributes combine the three privacy types with the 14 data categories. The additional privacy label is a *Data Not Collected* label with no associated data categories. The *Data Not Collected* label essentially offers a comparison to an app that has no privacy labels.

We completed two pilots with co-workers to refine the questions, and we also performed a final test run ($n = 20$) on *Prolific* [52]. Participants had to be 18 years of age or older and reside in the United States. There was no requirement for participants to be smartphone users. Below, we describe the final procedure in detail, and the complete survey instrument can be found in Appendix A.

1. *Informed Consent:* Participants consented to the study, risks, benefits, and right to withdraw.
2. *App Related Questions (Q1–Q7):* We presented each participant with the description (see the **Notebook** app in Appendix A) of a randomly assigned hypothetical iOS application and asked them to imagine they were making an install decision. We asked about participants' concern level and install history for the app type assigned to their study condition.
3. *Privacy Label Related Questions (Q8–Q12):* The image of a randomly selected Apple privacy label and questions about understanding, perceived risk, and willingness to install were displayed. Each participant was shown three labels and the same set of questions for each label.
4. *Demographics Questions (D1–D5):* Participants were asked (optionally) to provide demographic information, such as age, identified gender, and education.

Between-Subject Factor. We considered *app type* as a between-subject factor and tested two types of iOS apps: a social media app, which we hypothesized that most participants would have privacy concerns about [26, 34, 53]; and a note-taking app, which we expect to have fewer privacy concerns. To test this hypothesis, we asked participants how concerned they were about how the app would collect, store, and use their information and to explain their reasons. If they have this app installed on their device, we asked how long they have had it and why they installed it. If they did not have it installed, we inquired whether they had considered installing it and what deterred them from doing so.

Within-Subject Factor. In our study, we included 43 privacy label attributes. We tested three Apple privacy label privacy types, *Data Used to Track You*, *Data Linked to You*, and *Data Not Linked to You*, paired with one of 14 data categories, such as *Contact info*, *Location*, and *Purchases*. In addition, we included the *Data Not Collected* privacy label, which indicates that the app will not collect user data and does not pair with any of the 14 data categories. Out of the 43 privacy label attributes (shown in Table 3), each participant answered questions about three randomly selected privacy label attributes contextualized with a hypothetical app installation scenario. The implementation precludes a participant from being randomly assigned the same privacy label attribute multiple times. Each privacy label attribute is presented in the survey as an image of the privacy label as it would appear on the Apple app store when making an installation decision for an application. To evaluate how well participants believed they understood the label, we asked them how confident they knew what the presented label meant (**Q8**).

To gauge the participants' risk perception, we asked them to specify how the presented privacy label would change the privacy and security risks they associated with the specific app in question (**Q9-Q10**). Afterward, we asked participants to explain the reason behind their choice. We asked similar questions to ascertain the impact of the privacy label on changing participants' willingness to install the app (**Q11-Q12**).

Analysis Methods. We also used the same analysis methods as [23], including a large logistic regression with repeated measures to determine the likelihood of installing (or purchasing) an app and the associated risk perception. Notably, we utilized a repeated-measures design for the within-subject factor, in which we presented participants with similar question types in multiple scenarios. Consequently, three observations for each participant were not entirely independent. We accounted for this dependence using a statistical method that included random effects. Following [23] and prior work that modeled ordinal responses [7, 12, 24, 42, 59, 60, 63], we used Cumulative Link Mixed Models (CLMMs) with logit as the link function to assess the significance of our independent variables [15, 16]. The CLMM allowed us to model all five

levels of our ordinal Likert scale responses for our dependent variables: risk perception and willingness to install. We used a significance threshold of 0.05. We describe the methods in context for the remaining quantitative analysis.

For qualitative responses (five free-text questions), we utilize open coding to analyze the results of open-text questions. To achieve this, the research team's primary coder developed a codebook and identified descriptive themes for each question. Two secondary coders were responsible for coding a randomly sampled subset of 30% to ensure consistency and provided feedback on the codebook. Primary and secondary coders worked collaboratively to improve the codebook, iterating until inter-coder agreement was achieved (Cohen's $\kappa \geq 0.7$). Inter-rater reliability [50], measured with Cohen's κ ranged from 0.76 to 0.87 per question, with a mean of 0.80 (sd = 0.04). This level of agreement is "substantial" [43] or "excellent" [27].

We divided each qualitative response into two sets based on the app type assigned, note-taking, or social media. Due to the large number of responses, we used randomly sampled subsets of each free-text response. The size of the random subset (the percentage of responses for that particular question) was selected by the coders to reach thematic saturation, 20% for questions **Q2**, **Q10**, and **Q12**, 30% for **Q5**, and 65% for **Q7**.

Table 1: Demographic information of our study participants and the 2020 US Census data [10]. Categories not included in the US Census are denoted by –.

Metric	Levels	Study		Census
		n	**%**	**%**
Gender	Man	733	48.7	49.0
	Woman	724	48.1	51.0
	Non-binary	37	2.5	–
	Prefer not to disclose	10	0.7	–
	Prefer to self-describe	1	0.1	–
Age	18–24 years	324	21.5	12.9
	25–34 years	566	37.6	13.9
	35–44 years	341	22.7	12.7
	45–54 years	157	10.4	12.1
	55–64 years	83	5.5	13.0
	65+ years	31	2.1	16.8
	Prefer not to disclose	3	0.2	–
Education	No high school	23	1.5	13.9
	High school	156	10.4	26.6
	Some college	476	31.6	26.3
	Bachelor's degree	558	37.1	21.1
	Advanced degree	238	15.8	12.1
	Other	47	3.1	–
	Prefer not to disclose	7	0.5	–
Tech. Background	Yes	280	18.6	–
	No	1154	76.7	–
	Prefer not to disclose	71	4.7	–

Recruitment and Demographics. We recruited 1,505 participants via *Prolific* [52] for the survey between January 10,

2023 and January 20, 2023. Participants received $3.25 USD for completing the survey, and the median time to complete the survey was 8m, 41s. Participants were generally younger than the general population, with 21.5% between 18–24 years old, 37.6% between 25–34 years old, 22.7% between 35–44 years old, and 18% were 45 years or older. The identified gender distribution was 49% men, 48% women, and 3% non-binary, self-described, or chose not to disclose. Participant characteristics are presented in Table 1.

Limitations.   As an online survey, our ability to observe real app installations and ask follow-up questions to understand the full range of responses was limited. To compensate, we used thematic coding across multiple responses to capture opinions and feelings in a hypothetical installation scenario.

We also structured our study to measure the effectiveness of each privacy type and data category pair in isolation, allowing us to evaluate the impact of each label attribute and identify any misconceptions related to individual attributes. Nevertheless, as a complete privacy label would consist of more than one attribute, additional research is required to examine the subtleties in consumers' risk perception and willingness to install when presented with a complete Apple privacy label. It is expected that the impact of each attribute will be less pronounced when viewed in the context of a complete label, and interaction effects between label attributes may arise.

Some of our results may have been affected by social desirability bias, where participants overstated their privacy concerns or intention not to install an app. These results could be viewed as a potential upper bound on true behavior.

Furthermore, we acknowledge that our recruitment sample was younger and had higher educational attainment than the population overall (see Table 1). Still, our results offer valuable insights into willingness to install and risk perception upon viewing applications and associated privacy labels. Tang et al. [58] demonstrated that gender-balanced *Prolific* studies, including questions about user perceptions and experiences, provide reliable approximations of populations' behavior.

Ethical Considerations.   Our Institutional Review Board approved the study protocol. All participants were informed and consented to the study, and all collected data is only associated with random identifiers. We also reviewed each row in the dataset for potential personally identifiable information.

4   Results

This section is structured along our research questions. We first present our findings concerning participants' experiences and concerns with the apps they have installed or considered installing. Next, we show how participants understand the data collection information summarized in a privacy label. Finally,
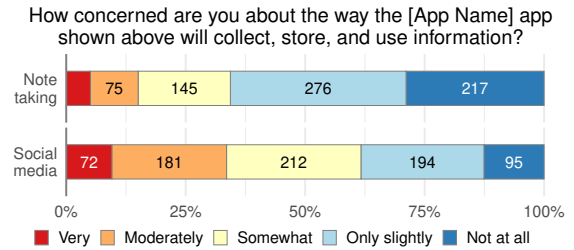
Figure 2: We asked participants to report their concern regarding the collection, storage, and use of information (**Q1**).

we discuss how privacy label attributes influence participants' risk perception and willingness to install an app.

4.1   RQ1: App Concern

In **RQ1**, we seek to understand participants' previous experiences installing social media and note-taking apps and their preexisting concerns regarding how those applications collect, store, and use their information.

App Concern Level   Participants were presented with a description of a randomly assigned generic note-taking app or a generic social media app. We then asked them to quantify their level of concern regarding the application's data collection and use (**Q1**). We hypothesized that participants assigned the social media app would report a greater level of concern than those assigned the note-taking app.

Quantitative. We found a strong correlation (Pearson's chi-square) between the app type and the level of concern. We considered the level of concern as a binary variable with *Not at all concerned* as 0, and all other concern levels as 1, $X^2(1, N = 1505) = 59.81$, $p < 0.001$, $\phi = 0.20$.

Of the participants who were assigned the note-taking app 34% reported being at least *Somewhat concerned* about how the app will collect, store, and use information. While 62% of participants assigned the social media app reported being at least *Somewhat concerned*. For full details regarding the levels of concern, please refer to Figure 2.

Qualitative. When describing their concern (**Q2**) for the note-taking application, common themes included concerns about their electronic notes being added to cloud storage or automatically synced across devices, the lack of information regarding data collection and use in the app description, and the potential for a data breach or exposure of their notes.

Common themes found when participants described their concern (**Q2**) for the social media app included unknown data collection and use policies, the reputation of social media apps for excessive data collection, sensitive information entered into the app, and data sold to third parties for targeted advertising. For instance, P805 (*Moderately concerned*) reported,

"The fact that it is free, and that social media companies are infamous for selling users' information."

Takeaway. The observed difference in participants' level of concern being greater for the social media app versus the note-taking app is consistent with our hypothesis and with previous research [26, 34, 53]. Emami-Naeini et al. [23] similarly found a strong correlation between the type of IoT device and participants' level of concern, with significant concern about smart speakers due to their always-listening capabilities.

Installation History. To understand participants' previous experiences installing an app of the type they were randomly assigned, we asked participants if they had installed an app of this type on their device (**Q3**), and if so, how long the app had been installed (**Q4**) and for what reason (**Q5**). For those who had not installed such an app, we asked if they had considered installing (**Q6**) and why they ultimately decided not to install (**Q7**). Responses to these questions allowed us to gain insight into participants' prior exposure to the apps and previous privacy concerns.

Quantitative. 88% reported having a social media app installed on their device, while only 49% reported installing a note-taking app (see Figure 5 in Appendix B). Among participants who had installed a social media app, 86% had the app for more than a year. For the note-taking app, 52% of participants indicated the app came preinstalled, and 33% had the app for over a year (see Figure 6 in Appendix B).

Among those who did not have a social media app installed, 76% reported that they had considered installing such an app. Of the participants who did not have a note-taking app installed, 38% reported that they had considered installing such an app on their device (refer to Figure 7 in Appendix B).

Qualitative. Participants most frequently reported (a) connecting with friends and family, (b) following and sharing content, (c) news and entertainment, (d) social pressure, and (e) accessing the social network on a mobile device, as reasons for installing a social media app. While the main reasons cited for installing a note-taking app included writing notes, making lists, keeping organized, setting reminders, syncing notes across devices, and storing important information.

The most common explanations for not installing a social media app included a dislike of social media, privacy concerns, too time-consuming or distracting, preferring to use a web browser to connect to the social media service, data collection concerns, mental health concerns, and concerns about sensitive data. Common reasons for not installing a note-taking app included that it did not meet current needs, that they would not use it often enough, preference for a physical notebook, and privacy concerns.

Takeaway. More participants reported they had previously installed a social media app than a note-taking app (88% to 49%). However, 52% of participants indicated that a note-taking app was preinstalled on their device. Privacy concerns were more often cited as the reason for not installing a social

media app than a note-taking app.

4.2 RQ2: Understanding of Privacy Labels

Apple's privacy labels provide considerable insights into the data collection practices of an application. With this research question, we measure participants' understanding of the information presented on the label. We evaluate a Likert question (**Q8**) about participants' confidence in the meaning of the information presented on the privacy label, as well as analyzing open-response questions (**Q10**, **Q12**) for misconceptions regarding the terminology used on the label.

Confidence Level in Understanding Label Information. For all but five privacy label data categories, more than 70% of participants reported (**Q8**) being *Somewhat confident*, *Moderately confident*, or *Very confident* about knowing what the privacy label information meant. However, for the data categories *Other data*, *Diagnostics*, *Identifiers*, *User content*, and *Sensitive info*, participants' level of confidence was significantly lower (p-value < 0.001). See Figure 12 in Appendix B for a full list of the data categories and a visualization of the responses. This result strongly corresponds to Emami-Naeini et al. [23], who found that over 70% of their study participants felt confident in understanding IoT privacy labels. Furthermore, like our study, they found that confidence was significantly lower for specific privacy label attributes, in their case, *security audit* and *data linkage*.

When considering privacy types, 96% of participants reported being at least *Somewhat confident* in their understanding of the label with a *Data Not Collected* privacy type. Participants reported being at least *Somewhat confident* only 73%, 71%, and 72% respectively for the *Data Used to Track You*, *Data Linked to You*, and *Data Not Linked to You* privacy types. Refer to Figure 8 in Appendix B for the full results.

We built a Cumulative Link Mixed Model (CLMM) to understand the impact of the privacy type and data category pairs on participants' confidence levels. We used the *Data Not Collected* privacy type as the baseline privacy label attribute. We found that when the data category *Other data* was paired with privacy types *Data used to track you*, *Data linked to you*, and *Data not linked to you* it was over 49 times, 103 times, and 43 times respectively more likely to cause a participant to reduce their confidence in understanding the label by one level. We also found that when the data category *Diagnostics* was paired with privacy types *Data used to track you*, *Data linked to you* , and *Data not linked to you* it was over 19 times, 27 times, and 20 times respectively more likely to cause a participant to reduce their confidence in understanding the label by one level. See Table 2 for the full CLMM results.

Takeaway. Participants reported confidence in understanding the privacy label, except for the data categories *Other data*, *Diagnostics*, *Identifiers*, *User content*, and *Sensitive info*.

Table 2: We used CLMM and built a model to identify the significance of various factors in changing participants' confidence in the meaning of the privacy label (**Q8**). For the 14 data categories the model captures the three privacy types for each category, i.e., *Data used to track you*, *Data linked to you*, and *Data not linked to you*.

Row	Factor	OR$_{(+)}$	OR$_{(-)}$	Estimate	Std. Error	p-value
				Confidence in meaning		
Data category by privacy type (baseline = Data not collected)						
1	Other data	0.02	48.86	-3.89	0.35	***
2	Diagnostics	0.05	19.27	-2.96	0.34	***
3	Identifiers	0.06	16.72	-2.82	0.33	***
4	User content	0.08	12.69	-2.54	0.32	***
5	Sensitive info	0.08	12.56	-2.53	0.33	***
6	Usage data	0.15	6.73	-1.91	0.32	***
7	Health & fitness	0.16	6.44	-1.86	0.33	***
8	Financial info	0.18	5.53	-1.71	0.33	***
9	Purchases	0.23	4.39	-1.48	0.33	***
10	Contact info	0.23	4.31	-1.46	0.33	***
11	Contacts	0.30	3.37	-1.22	0.34	***
12	Search history	0.37	2.69	-0.99	0.33	**
13	Location	0.67	1.49	-0.40	0.33	0.2
14	Browsing history	0.87	1.15	-0.14	0.35	0.7
15	Other data	0.01	103.00	-4.63	0.34	***
16	Diagnostics	0.04	26.90	-3.29	0.33	***
17	Identifiers	0.04	22.76	-3.12	0.32	***
18	User content	0.07	14.88	-2.70	0.31	***
19	Sensitive info	0.07	13.59	-2.61	0.31	***
20	Purchases	0.15	6.54	-1.88	0.32	***
21	Usage data	0.17	5.94	-1.78	0.31	***
22	Health & fitness	0.19	5.33	-1.67	0.32	***
23	Financial info	0.25	4.04	-1.40	0.32	***
24	Contact info	0.27	3.68	-1.30	0.32	***
25	Search history	0.40	2.51	-0.92	0.32	**
26	Contacts	0.54	1.84	-0.61	0.33	0.06
27	Browsing history	0.69	1.44	-0.37	0.33	0.27
28	Location	0.76	1.31	-0.27	0.32	0.40
29	Other data	0.02	43.41	-3.77	0.32	***
30	Identifiers	0.05	19.89	-2.99	0.31	***
31	User content	0.08	12.06	-2.49	0.31	***
32	Sensitive info	0.09	11.58	-2.45	0.30	***
33	Diagnostics	0.10	9.60	-2.26	0.31	***
34	Usage data	0.15	6.77	-1.91	0.30	***
35	Health & fitness	0.17	6.02	-1.79	0.31	***
36	Contact info	0.18	5.58	-1.72	0.31	***
37	Financial info	0.18	5.57	-1.72	0.31	***
38	Purchases	0.19	5.31	-1.67	0.31	***
39	Contacts	0.21	4.72	-1.55	0.31	***
40	Browsing history	0.29	3.51	-1.26	0.32	***
41	Location	0.30	3.30	-1.19	0.31	***
42	Search history	0.33	3.01	-1.10	0.31	***
Prior labels (baseline = 0 labels)						
43	1 label	0.88	1.14	-0.13	0.09	0.16
44	2 labels	0.89	1.13	-0.12	0.12	0.34
Threshold coefficients						
45	Not at all|Slightly	0.01	175.60	-5.17	0.28	***
46	Slightly|Somewhat	0.03	34.24	-3.53	0.27	***
47	Somewhat|Moderately	0.13	7.93	-2.07	0.26	***
48	Moderately|Very	0.90	1.11	-0.10	0.26	0.70
Random effects						
49	σ_u^2	–	–	2.86	–	–

Rows 1–14 grouped under *Data used to track you*; rows 15–28 under *Data linked to you*; rows 29–42 under *Data not linked to you*.

Note: *$p < 0.05$ **$p < 0.01$ ***$p < 0.001$

Common Misconceptions. Qualitative responses revealed common misunderstandings about the terminology used on the privacy label, including the terms track, linked, *Contact info*, and *Identifiers*, among others. Participants sometimes conflated the term tracking, which Apple defines as using data to track users across apps and websites owned by other companies, to mean tracking their interactions with the device. In a response about the *Diagnostics* data category with the *Data*

used to track you privacy type, P562 said, "I don't like the idea that they're tracking what sites I visit." Participants found the data collection associated with particular data categories to be unclear. For instance, P27 who was shown the *Identifiers* data category said, "Not sure what data is being collected." P292 confounded *Contact info* with their contacts when reporting, "I don't want my choices to potentially impact my contacts."

4.3 RQ3: Risk and Willingness to Install

To answer **RQ3**, we presented participants with a privacy label describing an app's data collection behavior. We then asked participants to rate the privacy and security risks on a Likert scale (**Q9**) and provide an open-ended explanation (**Q10**). Following this, we asked participants to rate, using a Likert scale (**Q11**) and an open-ended explanation (**Q12**), how the privacy label would impact their willingness to install the app.

CLMM Models. We developed two Cumulative Link Mixed Models (CLMMs) to assess how different factors influenced two dependent variables (DVs): participants' risk perception and willingness to install an iOS application (see Table 3). We included the following factors in each model:

- *Data category by privacy type*: 43 privacy label attributes consisting of three privacy types paired with the 14 data categories, and the *Data not collected* privacy type. Of the 43 attributes, only three were randomly chosen and shown to each participant, while the remaining attributes were not presented. We selected a label with the *Data not collected* privacy type as the baseline attribute as it is the one privacy type that has no associated data categories.
- *Label meaning confidence level*: The participant's confidence in the meaning of the label, with three levels: (a) *Not at all confident* or *Slightly confident*, (b) *Somewhat confident*, and (c) *Moderately confident* or *Very confident*. We used *Somewhat confident* as the baseline confidence level as it is the middle of the Likert values.
- *Application type*: We considered two levels of app type: social media and note-taking. The note-taking app was selected as the baseline because we expected its information use to be less concerning.
- *Concern about information use*: The participant's concern about the way the app will collect, store, and use information, with three levels: (a) *Not at all concerned* or *Slightly concerned*, (b) *Somewhat concerned*, and (c) *Moderately concerned* or *Very concerned*. We used *Somewhat concerned* as the baseline level of concern.
- *Prior labels*: Number of prior labels seen by that participant, with three levels: 0, 1, and 2 labels. We used zero prior labels as the baseline as it is the first level.
- *Participant age*: The age of the participant, with two levels: (a) less than 35 years old, and (b) 35 and older. We used 35 and older as the baseline age range because of its proximity to the median age of our participants.

Table 3: We used CLMM and built two models to identify the significance of various factors in changing participants' risk perception (**Q9**) and willingness to install (**Q11**). For the 14 data categories our models capture the three privacy types for each category, i.e., *Data used to track you*, *Data linked to you*, and *Data not linked to you*.

Row	Factor	Risk perception					Willingness to install				
		$OR_{(+)}$	$OR_{(-)}$	Estimate	Std. Error	*p*-value	$OR_{(+)}$	$OR_{(-)}$	Estimate	Std. Error	*p*-value
	Data category by privacy type (baseline = Data not collected)										
1	Financial info	11.43	0.09	2.44	0.33	***	0.00	641.94	-6.46	0.36	***
2	Sensitive info	10.27	0.10	2.33	0.32	***	0.00	303.76	-5.72	0.33	***
3	Other data	9.00	0.11	2.20	0.32	***	0.00	202.50	-5.31	0.33	***
4	Purchases	5.96	0.17	1.79	0.31	***	0.01	122.71	-4.81	0.32	***
5	Browsing history	5.95	0.17	1.78	0.32	***	0.01	151.97	-5.02	0.33	***
6	Contacts	5.91	0.17	1.78	0.32	***	0.01	161.22	-5.08	0.32	***
7	Search history	5.75	0.17	1.75	0.31	***	0.01	145.14	-4.98	0.32	***
8	Identifiers	5.70	0.18	1.74	0.32	***	0.01	113.27	-4.73	0.32	***
9	User content	4.41	0.23	1.48	0.31	***	0.01	94.38	-4.55	0.31	***
10	Contact info	4.19	0.24	1.43	0.32	***	0.01	108.36	-4.69	0.32	***
11	Location	4.18	0.24	1.43	0.31	***	0.01	90.83	-4.51	0.31	***
12	Health & fitness	3.98	0.25	1.38	0.31	***	0.01	79.93	-4.38	0.32	***
13	Usage data	3.55	0.28	1.27	0.30	***	0.02	54.33	-4.00	0.31	***
14	Diagnostics	3.05	0.33	1.11	0.31	***	0.02	41.82	-3.73	0.31	***
15	Financial info	14.40	0.07	2.67	0.32	***	0.00	363.33	-5.90	0.33	***
16	Sensitive info	9.80	0.10	2.28	0.31	***	0.00	406.68	-6.01	0.33	***
17	Browsing history	7.29	0.14	1.99	0.31	***	0.01	157.43	-5.06	0.31	***
18	Location	6.23	0.16	1.83	0.30	***	0.01	120.71	-4.79	0.30	***
19	Identifiers	5.51	0.18	1.71	0.31	***	0.01	102.92	-4.63	0.31	***
20	Search history	5.08	0.20	1.62	0.30	***	0.01	193.09	-5.26	0.31	***
21	Other data	4.66	0.21	1.54	0.31	***	0.01	106.21	-4.67	0.31	***
22	Contacts	4.11	0.24	1.41	0.31	***	0.01	144.23	-4.97	0.32	***
23	User content	4.00	0.25	1.39	0.30	***	0.01	98.56	-4.59	0.30	***
24	Contact info	3.95	0.25	1.37	0.30	***	0.02	65.25	-4.18	0.31	***
25	Usage data	3.51	0.28	1.26	0.30	***	0.03	34.65	-3.55	0.30	***
26	Diagnostics	3.30	0.30	1.19	0.30	***	0.04	23.17	-3.14	0.30	***
27	Health & fitness	3.19	0.31	1.16	0.30	***	0.02	50.73	-3.93	0.30	***
28	Purchases	3.15	0.32	1.15	0.30	***	0.01	75.80	-4.33	0.30	***
29	Financial info	5.71	0.17	1.74	0.30	***	0.01	95.54	-4.56	0.31	***
30	Sensitive info	3.64	0.27	1.29	0.29	***	0.03	37.47	-3.62	0.30	***
31	Identifiers	3.50	0.29	1.25	0.30	***	0.05	19.66	-2.98	0.30	***
32	Browsing history	3.22	0.31	1.17	0.30	***	0.03	38.80	-3.66	0.30	***
33	Location	3.20	0.31	1.16	0.30	***	0.04	27.09	-3.30	0.30	***
34	Search history	2.96	0.34	1.08	0.29	***	0.04	27.67	-3.32	0.30	***
35	Contact info	2.82	0.35	1.04	0.30	***	0.05	21.08	-3.05	0.30	***
36	Health & fitness	2.82	0.36	1.04	0.29	***	0.05	19.03	-2.95	0.29	***
37	Contacts	2.77	0.36	1.02	0.29	***	0.03	36.34	-3.59	0.30	***
38	Purchases	2.75	0.36	1.01	0.29	***	0.04	28.06	-3.33	0.29	***
39	Other data	2.59	0.39	0.95	0.30	**	0.05	19.52	-2.97	0.30	***
40	Usage data	2.51	0.40	0.92	0.28	**	0.09	11.30	-2.42	0.28	***
41	User content	2.41	0.41	0.88	0.29	**	0.05	22.00	-3.09	0.29	***
42	Diagnostics	2.09	0.48	0.74	0.29	*	0.09	11.49	-2.44	0.29	***
	Label meaning confidence (baseline = {*Somewhat*} confident)										
43	{*Very, Moderately*} confident	1.33	0.75	0.29	0.08	***	0.79	1.27	-0.24	0.08	**
44	{*Slightly, Not at all*} confident	1.04	0.96	0.04	0.09	0.67	0.57	1.77	-0.57	0.09	***
	App type (baseline = Note taking app)										
45	Social media app	0.79	1.27	-0.24	0.08	**	1.77	0.56	0.57	0.079	***
	Concern about app information use (baseline = {*Somewhat*} concerned)										
46	{*Very, Moderately*} concerned	1.05	0.95	0.05	0.11	0.64	0.70	1.43	-0.36	0.11	***
47	{*Slightly, Not at all*} concerned	0.81	1.23	-0.21	0.10	*	1.43	0.70	0.36	0.09	***
	Prior labels (baseline = 0 labels)										
48	1 label	0.86	1.16	-0.15	0.08	0.08	1.07	0.94	0.06	0.09	0.46
49	2 labels	0.79	1.27	-0.24	0.11	*	1.29	0.78	0.25	0.11	*
	Participant age (baseline = {*35 - 44, 45 - 54, 55 - 64, 65 or older*} years old)										
50	{*18-24, 25-34*} years old	1.03	0.97	0.03	0.08	0.71	1.70	0.59	0.53	0.08	***
	Threshold coefficients										
51	Strongly decreases | Slightly decreases	0.28	3.52	-1.26	0.27	***	0.01	105.60	-4.66	0.28	***
52	Slightly decreases | No impact	1.06	0.94	0.06	0.27	0.82	0.05	21.10	-3.05	0.27	***
53	No impact | Slightly increases	3.53	0.28	1.26	0.28	***	0.44	2.28	-0.83	0.27	**
54	Slightly increases | Strongly increases	17.27	0.06	2.85	0.28	***	3.42	0.29	1.23	0.26	***
	Random effects										
55	σ_u^2	–	–	1.12	–	–	–	–	0.82	–	–

Left margin labels: *Data used to track you* (rows 1–14), *Data linked to you* (rows 15–28), *Data not linked to you* (rows 29–42).

Note: *$p < 0.05$ **$p < 0.01$ ***$p < 0.001$

I believe the label shown above [Response] the privacy and security risks associated with this specific [App Name] app
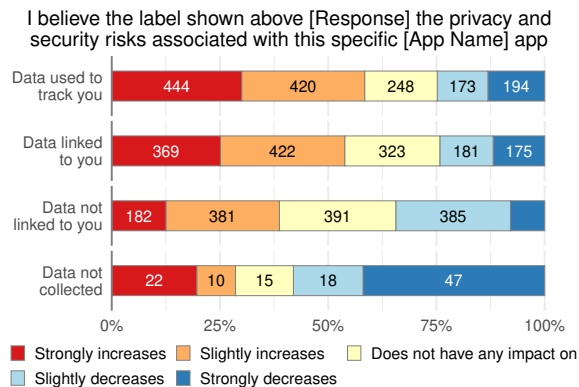
Figure 3: Participant risk perception by privacy type (**Q9**).

Assuming you want to install the [App Name] app on your phone, knowing that this app has the label shown above would [Response] your willingness to install this app

Figure 4: Participant willingness to install by label privacy type (**Q11**). *Data used to track you* caused the greatest decrease in willingness to install. While *Data not collected* increased participants' willingness to install.

We initially included additional factors such as whether the participant had installed the app, how long it had been installed, and if they considered installing the app. We also considered other demographic information, including gender, level of education, and technical background. However, the analysis revealed that these factors had little impact on the models, and so we removed these factors from our final models to improve goodness of fit, evaluated using the Akaike Information Criterion (AIC) [11]. Conversely, excluding other factors decreased model fit, so we retained all remaining factors in the models. Each model included a random intercept per participant to account for individual differences.

The CLMMs are trained on a dataset that includes three privacy label scenarios from each of the 1,505 participants, a total of 4,515 observations. Using a Likert scale, we asked participants to indicate the impact of the presented attribute, which comprised a privacy type and data category pair, on their risk perception and willingness to install the app (**Q9**, **Q11**).

In the risk perception model, a factor with a positive estimate suggests that risk perception has increased compared to the baseline for that factor. In the willingness to install model, a positive estimate indicates that participants are more inclined to install the iOS application. In contrast, a negative estimate suggests a reluctance to install compared to the baseline. In both models, all privacy type data category pairs on the privacy labels significantly affected participants' risk perception and willingness to install. Across all pairs, the impact was consistently in the direction of increased risk perception and decreased willingness to install. See Table 3 rows 1–42.

Risk Perception. According to the CLMM results, a label with the combination of *Financial info* with *Data linked to you* (Table 3, row 15) or *Financial info* with *Data used to track you* (Table 3, row 1) were the top two most significant impacts on increasing participants' risk perception. Additionally, the results indicate that a privacy label with the combination of *Sensitive info* with *Data used to track you* (Table 3, row 2) or the combination of *Sensitive info* with *Data linked to you*
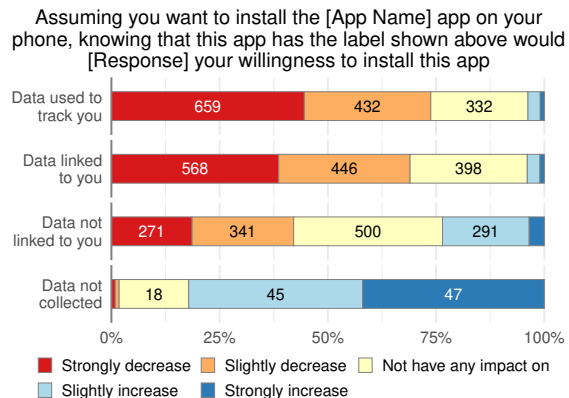
(Table 3, row 16) had the next two most significant impacts on increasing participants' risk perception.

The model results also show that the app type has a significant effect on the risk perception of the privacy label (Table 3, row 45). The participants assigned the social media app had a lower odds ratio, i.e., a *reduction* in privacy and security risk perception, than the participants assigned the note-taking app. This suggests that participants considered the label within the context of the app type and had a greater tolerance for or expected more data collection from a social media app.

Furthermore, the CLMM results show a decrease in the odds ratio for participants who reported being only *Slightly* or *Not at all* concerned about the way the app will collect, store, and use information (Table 3, row 47). It suggests that those participants who had less concern about the app's information use also had less concern about the data collection policy information displayed on the privacy label.

We observed a slight decrease in the odds ratios when the number of prior labels increased (Table 3, row 49), which suggests that viewing multiple privacy labels in a row for a single app causes a modest reduction in risk perception. This could be explained by participants being privacy resigned [21] or experiencing warning fatigue [14]. Participants could also be feeling lower risk compared to the previous label [23].

The CLMM results also found that the *Data not linked to you* privacy type (Table 3, rows 29–42) had lower odds ratios overall than the *Data used to track you* and *Data linked to you* privacy types (see Figure 3). Similarly, [23] found that attributes such as *data being sold to third parties* and *lack of access control* notably increased risk perception, while *no cloud retention* and *not sharing data with third parties* significantly reduced perceived risk.

Takeaway. All privacy label data categories increase risk perception. The data categories *Financial info*, *Sensitive info*,

and *Browsing History* were consistently the most likely to increase risk perception across privacy types. Participants with the social media app perceived lower risk than those with the note-taking app, contextualizing the risk by considering the app type. Participants who reported lower concern about app data collection also reported lower risk perception.

Willingness to Install.   The CLMM results showed that participants' willingness to install an app was significantly negatively impacted by privacy labels combining one of *Financial info* or *Sensitive info* with one of *Data linked to you* or *Data used to track you* (Table 3, rows 1, 2, 15, 16). This reduction in willingness to install aligns with the corresponding increase in risk perception for these same data categories.

The CLMM shows that when participants reported (**Q8**) being only *Slightly* or *Not at all* confident in the meaning of the privacy label, it had significant impacts on decreasing participants' willingness to install the app (Table 3, row 44). This shows that some participants would be reluctant to install an app whose privacy label they had difficulty understanding.

The model shows that the app type makes a significant impact (Table 3, row 45). The social media app *increased* participants' willingness to install the app, i.e., positive odds ratio. This suggests that participants consider the app type together with the data collection behavior disclosed on the privacy label when making an installation decision.

We found that participants who are under 35 played a significant positive factor in the willingness to install an app (Table 3, row 50), suggesting that younger users are more willing to install the applications regardless of privacy labels.

We also found that the *Data not linked to you* privacy type (Table 3, rows 29–42) had higher odds ratios (increase in willingness to install) overall than both the *Data used to track you* and *Data linked to you* privacy types. Figure 4 shows the full results of the willingness to install by privacy type.

Comparing the odds ratios of risk perception and willingness to install (presented in Table 3), we observe that for all of the label attributes, the odds ratios of decreasing willingness to install are higher than their corresponding odds ratios of increasing risk perception. This finding suggests that the tested label attributes had a greater impact on participants' willingness to install an app than on their risk perception.

Takeaway. The CLMM showed that participants were less willing to install apps when the labels showed that *Financial info* or *Sensitive info* was tracked or linked to them. Those uncertain about the meaning of privacy labels were less inclined to install the app. Younger participants (under 35) were more lenient regarding installation. Overall, privacy labels significantly influenced installation decisions, *more* than they impacted risk perception. This contradicts [23], which found that labels were *less* influential in altering willingness to purchase an *IoT device* than in altering risk perception.

Response Category Analysis.   Based on the CLMM estimates, we computed the probabilities of the five response categories for risk perception and willingness to install (Appendix B Figure 11). Participants were more likely to express increased risk perception for all label attributes except *Data not collected*. For most data categories, *Data used to track you* correlated with the highest probability for increased risk perception. However, *Data linked to you* had the highest correlation for *Browser history*, *Financial info*, and *Location*.

For most data categories, when combined with *Data used to track you* or *Data linked to you*, the highest probable response was a *Strong decrease* in their willingness to install the app. The *Diagnostics* and *Usage data* were exceptions, where the responses with the highest probability were *Slightly decrease your willingness to install* or *Not have any impact on your willingness to install* was the highest. This suggests that participants were more accepting of data collection if it was associated with improving the application or if they found the information collected less sensitive.

The *Data linked to you* privacy type had higher probabilities in the *Strongly decrease your willingness to install* response on 9 of the 14 data categories. The *Data linked to you* privacy type played a more prominent role in the reduction of participants' willingness to install an app than in risk perception, where the *Data used to track you* privacy type was the leader in increasing a participants' privacy and security risks. This suggests that tracking collected data is more highly associated with privacy and security risks to participants, whereas linking data is more of a deterrent to installing an application.

Reasons for Concern.   Participants' replies to the open-ended questions provide a deeper understanding of the reasoning behind risk perceptions and willingness to install an app. Participants reported concern that the collected data was personally identifiable. They were also concerned about the collection of private or sensitive information, tracking, and unauthorized access (e.g., in a breach or through misuse). When presented with the *Identifiers* data category combined with the *Data linked to you* privacy type, P246 expressed concern that the collection was personally identifiable: "This sounds like it would be information that could be specifically linked to me and me alone." P77 responded (social media app, *Health & fitness*, *Data linked to you*), "I don't want them to know my health info." P1481 shared concerns regarding unauthorized access: "Storing personal and sensitive information in a place the user is unaware of and in a system that could be hacked could mean that information could get into the wrong hands and it's completely outside of the user's control." P1030 (social media app, *Sensitive info*, *Data used to track you*) stated, "I would not want any sensitive information shared, so I would not install an app with this label."

Participants also had common reasons for reduced concern, such as data not being linked to their identity, data not collected, limited data collection, and data categories they did not

consider sensitive. P26 shared that the *Location* data category combined with *Data not linked to you* privacy type *does not have any impact on the privacy and security risks* "because the location is not linked to my identity." P1142 said that the *Data not collected* privacy type, "means that the app developer does not collect any data, so therefore there is no risk to privacy and security because they have no information about you." P1244 (*User content*, *Data linked to you*) found the data collection to be limited, "It shows a level of transparency and also that they are taking labels seriously by doing the minimum." And P956, who was not concerned about the sensitivity of the *Contact info* data category, said, "My contact information is less of a concern than more personal data such as financial information the app may need from me."

Privacy Resigned. Some participants are resigned to data collection as the new standard for applications. P8 (social media app, *Contacts*, *Data used to track you*, *No impact*) shared this example, "Most apps already track all my information and location, so I would not be worried about one more having it." And P1091 (note-taking app, *Health & fitness*, *Data not linked to you*, *No impact*) replied, "I assume all apps and Apple products talk to one another and spy on me." While P77 (social media app, *Identifiers*, *Data used to track you*, *No impact*) added, "Other apps and companies track me on websites already." Participant P808 (social media app, *Health & fitness*, *Data used to track you*, *No impact*) simply said, "There are always privacy risks when using any kind of app."

Lack of Transparency. Some participants complained that the first level privacy label, i.e., privacy types and data categories, did not provide enough transparency about data collection and application practices to make an informed decision regarding their willingness to install the app. Many still wanted to know how their collected information would be used, why the data is collected, who would have access, precisely what data is collected, and how it is protected once obtained. Participant P515 (social media app, *Contacts*, *Data not linked to you*) said, for example, "It doesn't necessarily reassure me about how my data will be used." While P854 (note-taking app, *Other data*, *Data used to track you*) had concerns about how the data is used and shared, "It allows data to be shared to other companies and does not specify exact what it would be used for or how safe it will be." And P1254 (note-taking app, *Identifiers*, *Data linked to you*), who was concerned that the label, while reporting the data collected, did not give any indication about the risks that are incumbent part of that data collection added, "It is merely showing what is stored as data not the risks associated with it."

Lack of Trust. Some participants did not trust the app developers to adhere to the practices reported in the privacy label. For instance, P1076 (note-taking app, *Other data*) said

of the *Data not linked to you* privacy type, "There is less worry about information going in the cloud because it supposedly cannot be linked back to me, but I'm still not 100% convinced that any online data can be completely un-linkable to you." P535 (social media app, *User content*, *Data not linked to you* privacy type) replied, "I feel that the wording is not that trustworthy and I feel that some data will still be linked to me in some way." P597 (note-taking app, *Usage data*, *Data not linked to you*) shared, "Software companies routinely claim privacy but have proven to be false." P1014 did not trust the *Data not collected*: "I don't believe I would trust that this claim is true." P179 who also viewed the *Data not collected* label simply added, "I assume it is lying to some extent." A lack of trust can undermine the usefulness of privacy labels.

Privacy Tradeoffs. Some participants expressed the need to trade their privacy through the data collected for the app's utility or the fact that it might be free. For example, when P956 (social media app, *Search history*, *Data used to track you*) explained, "It may be that my desire to have the app outweighs the risk until something actually happens." Furthermore, P1212, assigned the social media app, said, "Most social media apps collect data. I've come to expect it from free ones because I know they need to make money." And P662 (social media app, *Search history*, *Data used to track you*) added, "It is a small price to pay for free apps."

5 Discussion and Conclusion

Privacy Labels Across Contexts As our study replicated the methodology of an IoT privacy label study [23], we can compare the results of both studies to understand how privacy labels function in different contexts. In both studies, over 70% of participants felt confident in their understanding of the labels presented, which is a promising result for the usability of privacy labels. However, both studies found that label attributes that included technical jargon caused participant understanding to be significantly lowered, such as *security audit* and *data linkage* for IoT labels and *Other data* and *Diagnostics* for iOS labels. This result suggests that privacy label attributes should be free of technical jargon and use terminology comprehensible to a broad audience.

Both studies observed differences in participants' level of concern regarding the type of app or product under consideration. In the IoT study, there was more concern about smart speakers due to their always-listening capabilities. In our iOS study, there was more concern regarding social media apps due to their reputation for excessive data collection.

Moreover, both studies found that privacy label attributes involving tracking, linking, or selling consumer data to third parties significantly increased participants' risk perception. Furthermore, all privacy label attributes that reduced consumer data protection in the IoT study and all privacy la-

bel attributes except the *Data not collected* attribute in our iOS study increased participants' risk perception. The results demonstrate that labels can effectively be used as a privacy disclosure mechanism in a variety of contexts to communicate the risks of personal data collection, storage, and use.

Transparency Paradox   Our qualitative responses show participants complaining about the vagueness of privacy labels and the lack of transparency. Participants found that the label did not provide the level of detail necessary to determine whether an app's data collection increases the security and privacy risks associated with its use. Thus, they found it difficult to decide whether or not to install the application. This leads us to believe that privacy labels suffer from the transparency paradox, the inherent conflict between transparency of textual meaning and the transparency of privacy practices [51]. Summarizing information handling practices in the form of privacy nutrition labels removes relevant details needed for people to make meaningful choices regarding their privacy. This loss of informational complexity, in turn, leads to a loss in specificity. Reducing informational complexity is a laudable goal; however, it is important to recognize that excessive summarizing of privacy information may lead to confusion and mistrust, especially among users who want to fully understand the implications of the data collection. Participants felt that the first level privacy label did not specify how their data would be used, why it was collected, who would have access, and how it would be protected. However, providing too much detail, such as through a privacy policy, can overwhelm users and deter them from reading privacy information. Prior work has suggested the use of hover text [62] or providing an info link to offer another layer of explanation. Further research can help us find a balance between granularity and effectiveness.

Balancing Comfort with Complacency   Our qualitative responses revealed that people are more willing to trust an application when the privacy label provides information about what data will be collected. Providing this information upfront reassures users that the developer is not trying to collect data without their consent or knowledge. However, this can lead to complacency if users do not additionally determine whether the collected data is necessary. This suggests that privacy labels might give consumers a false sense of security, leading them to believe that data collection cannot be harmful if informed about it. This raises the critical question: Do these labels create comfort for consumers but fail to provide actual privacy? Trust in this context could be harmful if it leads to complacency or disinformation. Prior work has shown that privacy labels are often inaccurate due to a lack of oversight and developer confusion when creating labels [45]. One possible solution is establishing more effective oversight mechanisms to ensure that privacy labels are accurate and truthful. Additionally, more education is needed to help consumers understand the limitations of privacy labels and

encourage them to take a more active role in protecting their data. It is crucial to balance transparency and accountability to promote informed decision-making and protect privacy.

Impact of Alternatives on Willingness to Install   While consumers can find alternatives for certain apps (e.g., note-taking), others (e.g., social media) are harder to replace. Consumers' willingness to install an app that collects data they are uncomfortable sharing depends on the app's necessity and their willingness to make privacy tradeoffs. Labels provide consumers with an easy way to comparison shop for apps that align with their preferences, assuming the app fulfills their requirements. With over 1.5 million apps available, consumers can choose from multiple alternatives. However, with limited choices, users may feel forced to make a privacy tradeoff. Emami-Naeini et al. [23] found that, in a marketplace with few alternatives, labels were more influential in changing risk perception than in altering willingness to purchase, suggesting that while privacy and security are important factors, they are among several factors, including price, features, and quality, that are considered by consumers when deciding to purchase an IoT device. Our study suggests that the availability of suitable alternatives can impact users' willingness to install an app due to privacy concerns.

Data Collection in Context   We found that participants were savvy when matching the category of the collected data within the context of the app. For instance, they reported being wary of the note-taking app collecting *Financial info* or *Location* data since the data seemed out of alignment with the app type. The label made them question the motives for collecting data not needed for the application's functionality, and when it does not make sense in context, the practice reduces trust in the app and the developer. Further research can study additional app contexts with label information.

Impact of Privacy Labels   Our study found that labels significantly impact users' risk perception and willingness to install an app. Accurate labels have the potential to communicate risk and help consumers align their privacy expectations with real-world privacy outcomes even more than other disclosure mechanisms (e.g., privacy policies). Even when participants reported limited understanding of certain attributes (e.g., *Other data*, the labels made them question associated risks. Discomfort with unknowns, such as what data is collected, emerged as a common theme in our qualitative responses. These findings underscore the importance of labels to empower consumers to make informed decisions. However, as our study also shows, the effectiveness of these labels is contingent on their accuracy and comprehensiveness. Further research is necessary to understand how to optimize the design and implementation of privacy labels to better serve consumers and promote a more transparent app ecosystem.

References

[1] How to Use Apple's Privacy Labels for Apps, 2023. URL: https://www.consumerreports.org/privacy/how-to-use-apples-privacy-labels-for-apps-a1059836329/.

[2] Paarijaat Aditya, Bobby Bhattacharjee, Peter Druschel, Viktor Erdélyi, and Matthew Lentz. Brave new world: Privacy risks for mobile users. In *Proceedings of the ACM MobiCom workshop on Security and privacy in mobile environments*, pages 7–12, 2014.

[3] Federal Trade Commission: Consumer Advice. How To Use the EnergyGuide Label To Shop for Home Appliances. http://consumer.ftc.gov/articles/how-use-energyguide-label-shop-home-appliances, May 2021. URL: http://consumer.ftc.gov/articles/how-use-energyguide-label-shop-home-appliances.

[4] Hazim Almuhimedi, Florian Schaub, Norman Sadeh, Idris Adjerid, Alessandro Acquisti, Joshua Gluck, Lorrie Faith Cranor, and Yuvraj Agarwal. Your Location has been Shared 5,398 Times! A Field Study on Mobile App Privacy Nudging. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 787–796, New York, NY, USA, April 2015. Association for Computing Machinery. doi:10.1145/2702123.2702210.

[5] Apple. Apple AppStore, May 2022. publisher: Apple. URL: https://apps.apple.com/.

[6] Apple. App Store Review Guidelines - Apple Developer. https://developer.apple.com/app-store/review/guidelines, 2023. Last Accessed: April 26, 2023.

[7] Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A. Hearst, Andrew Head, and Kyle Lo. Paper plain: Making medical research papers approachable to healthcare consumers with natural language processing. *ACM Trans. Comput.-Hum. Interact.*, 30(5), sep 2023. doi:10.1145/3589955.

[8] David G Balash, Mir Masood Ali, Xiaoyuan Wu, Chris Kanich, and Adam J Aviv. Longitudinal analysis of privacy labels in the apple app store. *arXiv preprint arXiv:2206.02658*, 2022.

[9] Rebecca Balebako, Florian Schaub, Idris Adjerid, Alessandro Acquisti, and Lorrie Cranor. The Impact of Timing on the Salience of Smartphone App Privacy Notices. In *Proceedings of the 5th Annual ACM CCS Workshop on Security and Privacy in Smartphones and Mobile Devices*, SPSM '15, page 63–74, New York, NY, USA, 2015. Association for Computing Machinery. doi:10.1145/2808117.2808119.

[10] US Census Bureau. Age and Sex Composition in the United States: 2020. https://www.census.gov/data/tables/2020/demo/age-and-sex/2020-age-sex-composition.html, 2020.

[11] Kenneth P Burnham and David R Anderson. Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research*, 33(2):261–304, 2004.

[12] Zekun Cai and Aiping Xiong. Understand users' privacy perception and decision of V2X communication in connected autonomous vehicles. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 2975–2992, Anaheim, CA, August 2023. USENIX Association. URL: https://www.usenix.org/conference/usenixsecurity23/presentation/cai-zekun.

[13] Brian X. Chen. What We Learned From Apple's New Privacy Labels. 2021. URL: https://www.nytimes.com/2021/01/27/technology/personaltech/apple-privacy-labels.html.

[14] Hanbyul Choi, Jonghwa Park, and Yoonhyuk Jung. The role of privacy fatigue in online privacy behavior. *Computers in Human Behavior*, 81:42–51, 2018.

[15] Rune Haubo B Christensen. Cumulative link models for ordinal regression with the r package ordinal. *Submitted in J. Stat. Software*, 35, 2018.

[16] Rune Haubo B Christensen. A Tutorial on fitting Cumulative Link Mixed Models with clmm2 from the ordinal Package. https://cran.ms.unimelb.edu.au/web/packages/ordinal/vignettes/clmm2_tutorial.pdf, 2019.

[17] Karen Church, Denzil Ferreira, Nikola Banovic, and Kent Lyons. Understanding the challenges of mobile phone usage data. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 504–514, 2015.

[18] European Commission. Energy-efficient products. https://ec.europa.eu/info/energy-climate-change-environment/standards-tools-and-labels/products-labelling-rules-and-requirements/energy-label-and-ecodesign/energy-efficient-products_en, 2022. Last Accessed: October 28, 2022.

[19] Lorrie Faith Cranor. Necessary But Not Sufficient: Standardized Mechanisms for Privacy Notice and Choice. *J. on Telecomm. & High Tech. L.*, 10:273, 2012. URL: http://jthtl.org/content/articles/V10I2/JTHTLv10i2_Cranor.PDF.

[20] Lorrie Faith Cranor, Candice Hoke, Pedro Leon, and Alyssa Au. Are They Worth Reading? An In-Depth Analysis of Online Advertising Companies' Privacy Policies. SSRN Scholarly Paper ID 2418590, Social Science Research Network, Rochester, NY, March 2014. URL: `https://papers.ssrn.com/abstract=24185 90`.

[21] Nora A. Draper. From privacy pragmatist to privacy resigned: Challenging narratives of rational choice in digital privacy debates. *Policy & Internet*, 9(2):232–251, 2017.

[22] Pardis Emami-Naeini, Yuvraj Agarwal, Lorrie Faith Cranor, and Hanan Hibshi. Ask the Experts: What Should Be on an IoT Privacy and Security Label? In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 447–464, San Jose, CA, USA, May 2020. IEEE. ISSN: 2375-1207. `doi:10.1109/SP40000.2020.00043`.

[23] Pardis Emami-Naeini, Janarth Dheenadhayalan, Yuvraj Agarwal, and Lorrie Faith Cranor. Which Privacy and Security Attributes Most Impact Consumers' Risk Perception and Willingness to Purchase IoT Devices? In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 519–536, San Francisco, CA, USA, May 2021. IEEE. URL: `https://ieeexplore.ieee.org/document/9 519463/`, `doi:10.1109/SP40001.2021.00112`.

[24] Pardis Emami-Naeini, Janarth Dheenadhayalan, Yuvraj Agarwal, and Lorrie Faith Cranor. Are consumers willing to pay for security and privacy of IoT devices? In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 1505–1522, Anaheim, CA, August 2023. USENIX Association. URL: `https://www.usenix.o rg/conference/usenixsecurity23/presentatio n/emami-naeini`.

[25] Adrienne Porter Felt, Serge Egelman, and David Wagner. I've got 99 problems, but vibration ain't one: a survey of smartphone users' concerns. In *Proceedings of the second ACM workshop on Security and privacy in smartphones and mobile devices*, pages 33–44, 2012.

[26] Elizabeth Fife and Juan Orjuela. The privacy calculus: Mobile apps and user perceptions of privacy and security. *International Journal of Engineering Business Management*, 4(Godište 2012):4–11, 2012.

[27] Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. *Statistical methods for rates and proportions*. John Wiley & Sons, Hoboken, New Jersey, 2013.

[28] FDA Center for Devices and Radiological Health. Device Labeling. `https://www.fda.gov/medical-dev ices/overview-device-regulation/device-lab eling`, October 2020. URL: `https://www.fda.gov/ medical-devices/overview-device-regulatio n/device-labeling`.

[29] Marco Furini, Silvia Mirri, Manuela Montangero, and Catia Prandi. Privacy perception when using smartphone applications. *Mobile Networks and Applications*, 25:1055–1061, 2020.

[30] Jack Gardner, Yuanyuan Feng, Kayla Reiman, Zhi Lin, Akshath Jain, and Norman Sadeh. Helping mobile application developers create accurate privacy labels. In *2022 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 212–230, 2022. `doi:10.1109/EuroSPW55150.2022.00028`.

[31] Rajiv Garg and Rahul Telang. Impact of app privacy label disclosure on demand: An empirical analysis. *Workshop on the Economics of Information Security (WEIS)*, 2022.

[32] Apple Inc. App Privacy Details - App Store, 2020. URL: `https://developer.apple.com/app-store/app -privacy-details/`.

[33] Carlos Jensen and Colin Potts. Privacy Policies as Decision-Making Tools: An Evaluation of Online Privacy Notices. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 471–478, 2004.

[34] Mohsen Jozani, Emmanuel Ayaburi, Myung Ko, and Kim-Kwang Raymond Choo. Privacy concerns and benefits of engagement with social media-enabled apps: A privacy calculus perspective. *Computers in Human Behavior*, 107:106260, 2020.

[35] Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W. Reeder. A "Nutrition Label" for Privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, SOUPS '09, pages 1–12, New York, NY, USA, July 2009. Association for Computing Machinery. `doi:10.1145/1572532.1572538`.

[36] Patrick Gage Kelley, Lucian Cesca, Joanna Bresee, and Lorrie Faith Cranor. Standardizing Privacy Notices: An Online Study of the Nutrition Label Approach. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1573–1582, New York, NY, USA, April 2010. Association for Computing Machinery. `doi:10.1145/1753326.1753561`.

[37] Patrick Gage Kelley, Sunny Consolvo, Lorrie Faith Cranor, Jaeyeon Jung, Norman Sadeh, and David Wetherall. A conundrum of permissions: installing applications on an android smartphone. In *Financial Cryptography and Data Security: FC 2012 Workshops, USEC and WECSR 2012, Kralendijk, Bonaire, March 2, 2012, Revised Selected Papers 16*, pages 68–79. Springer, 2012.

[38] Patrick Gage Kelley, Lorrie Faith Cranor, and Norman Sadeh. Privacy as part of the app decision-making process. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3393–3402, Paris France, April 2013. ACM. URL: `https://dl.acm.org/doi/10.1145/2470654.2466466`, `doi:10.1145/2470654.2466466`.

[39] Hammad Khalid, Emad Shihab, Meiyappan Nagappan, and Ahmed E Hassan. What do mobile app users complain about? *IEEE software*, 32(3):70–77, 2014.

[40] Jennifer King. How come i'm allowing strangers to go through my phone? smartphones and privacy expectations. *Smartphones and Privacy Expectations.(March 15, 2012)*, 2012.

[41] Konrad Kollnig, Anastasia Shuba, Max Van Kleek, Reuben Binns, and Nigel Shadbolt. Goodbye Tracking? Impact of iOS App Tracking Transparency and Privacy Labels. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, Virtual Conference, April 2022. Association for Computing Machinery. arXiv: 2204.03556. URL: `http://arxiv.org/abs/2204.03556`, `doi:10.1145/3531146.3533116`.

[42] Marvin Kowalewski, Christine Utz, Martin Degeling, Theodor Schnitzler, Franziska Herbert, Leonie Schaewitz, Florian M. Farke, Steffen Becker, and Markus Dürmuth. 52 weeks later: Attitudes towards covid-19 apps for different purposes over time. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2), oct 2023. `doi:10.1145/3610042`.

[43] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, 33(1):159–174, 1977.

[44] Nicholas D Lane, Emiliano Miluzzo, Hong Lu, Daniel Peebles, Tanzeem Choudhury, and Andrew T Campbell. A survey of mobile phone sensing. *IEEE Communications magazine*, 48(9):140–150, 2010.

[45] Tianshi Li, Kayla Reiman, Yuvraj Agarwal, Lorrie Faith Cranor, and Jason I. Hong. Understanding Challenges for Developers to Create Accurate Privacy Nutrition Labels. In *CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. `doi:10.1145/3491102.3502012`.

[46] Tong Li, Yong Li, Tong Xia, and Pan Hui. Finding spatiotemporal patterns of mobile application usage. *IEEE Transactions on Network Science and Engineering*, 2021.

[47] Tong Li, Tong Xia, Huandong Wang, Zhen Tu, Sasu Tarkoma, Zhu Han, and Pan Hui. Smartphone app usage analysis: datasets, methods, and applications. *IEEE Communications Surveys & Tutorials*, 2022.

[48] Jialiu Lin, Shahriyar Amini, Jason I Hong, Norman Sadeh, Janne Lindqvist, and Joy Zhang. Expectation and purpose: understanding users' mental models of mobile app privacy through crowdsourcing. In *Proceedings of the 2012 ACM conference on ubiquitous computing*, pages 501–510, 2012.

[49] Aleecia M. McDonald and Lorrie Faith Cranor. The Cost of Reading Privacy Policies. *HeinOnline*, 4(3):543–568, 2009. URL: `https://heinonline.org/HOL/P?h=hein.journals/isjlpsoc4&i=563`.

[50] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), nov 2019. `doi:10.1145/3359174`.

[51] Helen Nissenbaum. A Contextual Approach to Privacy Online. *Daedalus*, 140(4):32–48, 2011.

[52] Prolific, Academic Ltd. A Higher Standard of Online Research, December 2022. `https://www.prolific.co`, as of June 10, 2024.

[53] Li Qin, Yongbeom Kim, and Xin Tan. Understanding the intention of using mobile social networking apps across cultures. *International Journal of Human–Computer Interaction*, 34(12):1183–1193, 2018.

[54] Joel R Reidenberg, Travis Breaux, Lorrie Faith Cranor, Brian French, Amanda Grannis, James T Graves, Fei Liu, Aleecia McDonald, Thomas B Norton, and Rohan Ramanath. Disagreeable Privacy Policies: Mismatches between Meaning and Users' Understanding. *Berkeley Tech. LJ*, 30:39, 2015.

[55] Florian Schaub, Rebecca Balebako, Adam L. Durity, and Lorrie Faith Cranor. A Design Space for Effective Privacy Notices. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 1–17, Ottawa, Canada, July 2015. USENIX Association. URL: `https://www.usenix.org/conference/soups2015/proceedings/presentation/schaub`.

[56] Irina Shklovski, Scott D Mainwaring, Halla Hrund Skúladóttir, and Höskuldur Borgthorsson. Leakiness and creepiness in app space: Perceptions of privacy and mobile app use. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2347–2356, 2014.

[57] FTC Staff. Protecting consumer privacy in an era of rapid change–a proposed framework for businesses and policymakers. *Journal of Privacy and Confidentiality*, 3(1), 2011. URL: `https://www.ftc.gov/reports/protecting-consumer-privacy-era-rapid-change-recommendations-businesses-policymakers`.

[58] Jenny Tang, Eleanor Birrell, and Ada Lerner. Replication: How Well Do My Results Generalize Now? The External Validity of Online Privacy and Security Surveys. In *Proc. SOUPS '22*, pages 367–385, Boston, Massachusetts, USA, August 2022. USENIX Association.

[59] Christine Utz, Steffen Becker, Theodor Schnitzler, Florian M. Farke, Franziska Herbert, Leonie Schaewitz, Martin Degeling, and Markus Dürmuth. Apps against the spread: Privacy implications and user acceptance of covid-19-related smartphone apps on three continents. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. `doi:10.1145/3411764.3445517`.

[60] Zixin Wang, Danny Yuxing Huang, and Yaxing Yao. Exploring tenants' preferences of privacy negotiation in airbnb. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 535–551, Anaheim, CA, August 2023. USENIX Association. URL: `https://www.usenix.org/conference/usenixsecurity23/presentation/wang-zixin`.

[61] Yue Xiao, Zhengyi Li, Yue Qin, Xiaolong Bai, Jiale Guan, Xiaojing Liao, and Luyi Xing. Lalaine: Measuring and characterizing Non-Compliance of apple privacy labels. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 1091–1108, Anaheim, CA, August 2023. USENIX Association. URL: `https://www.usenix.org/conference/usenixsecurity23/presentation/xiao-yue`.

[62] Shikun Zhang, Yuanyuan Feng, Yaxing Yao, Lorrie Faith Cranor, and Norman Sadeh. How Usable Are iOS App Privacy Labels? *Proceedings on Privacy Enhancing Technologies*, 4:204–228, 2022.

[63] Yuhang Zhao, Yaxing Yao, Jiaru Fu, and Nihan Zhou. "If sighted people know, i should be able to know:" privacy perceptions of bystanders with visual impairments around camera-based technology. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 4661–4678, Anaheim, CA, August 2023. USENIX Association. URL: `https://www.usenix.org/conference/usenixsecurity23/presentation/zhao-yuhang`.

A Survey Instrument

Thank you for your interest in our survey.
Please read the following instructions carefully: (i) Take your time in reading and answering the questions. (ii) Answer the questions as accurately as possible.
Definitions: (i) App: In this survey the word "app" refers to an application found on the Apple App Store that can be installed on your Apple device. (ii) Privacy Label: a short summary of an app's data collection behavior displayed on the application pages of the Apple App Store.
On the next page we will provide an introduction to this survey.
[A horizontal rule, like below, indicates a new page in the questionnaire.]

Survey Introduction

This survey is designed to investigate your awareness of app privacy labels displayed on the application pages of the Apple App Store. You will answer questions regarding potential app installation decisions and how an app privacy label may impact your thoughts about the app.
On the following pages you will be presented with an application and asked questions about this application and its privacy labels. For each of the labels we will ask a set of similar questions, so please pay close attention.

App Related Questions

[Apps are randomly assigned.]
Imagine you are making a decision to install a *[App Name]* app on your phone that was recommended by a friend. The price of the app is within your budget (or it is free) and the features are what you would expect from a *[App Name]* app.
Assume you do not have a *[App Name]* app installed on your phone.
Please review the app description before answering the questions.
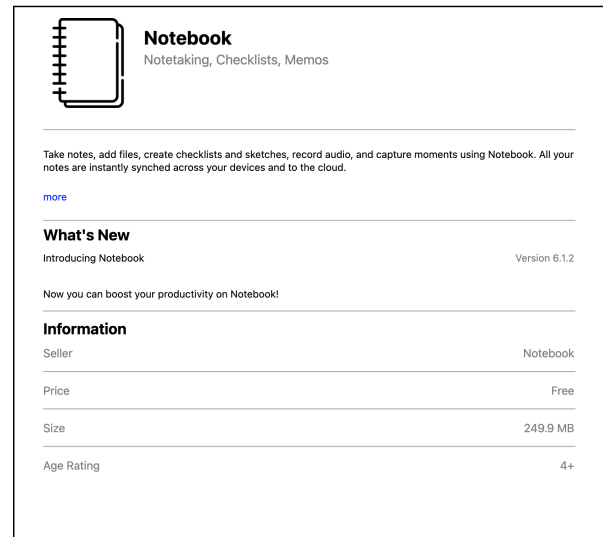
[An example image of a note taking app displayed to participants.]

Q1 How concerned are you about the way the *[App Name]* app shown above will collect, store, and use information?
- ◯ Not at all concerned
- ◯ Slightly concerned
- ◯ Somewhat concerned
- ◯ Moderately concerned
- ◯ Very concerned

Q2 What about data collection, storage, and use by the *[App Name]* app makes you feel concerned?
- Please type in your answer.

Q3 Do you currently have a *[App Name]* app installed on your phone?

○ Yes   ○ No

[Included only if Yes selected in Q3.]

Q4 How long have you had this *[App Name]* app? If you have more than one device, answer the question for the one that you have had for the longest time.
○ Less than a month   ○ More than a year
○ Between a month and a year   ○ I don't remember

Q5 What were your reasons to install the *[App Name]* app?
- Please type in your answer.

[Included only if No selected in Q3.]

Q6 Have you ever considered installing a *[App Name]* app on your phone?

○ Yes   ○ No

[Included only if Yes selected in Q6.]

Q7 What made you decide not to install the *[App Name]* app?
- Please type in your answer.

Privacy Label Related Questions

[Q8 - Q12 will be asked once per privacy label. The privacy labels are chosen randomly. Participants are shown and asked to respond to three privacy labels.]

Please imagine the following privacy label (a short summary of the app's data collection behavior) was shown on the App Store page of the app when answering the questions below.

Data Linked to You

The following data may be collected and linked to your identity:

🛍 Purchases
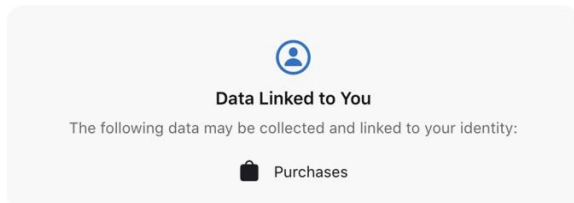
[An example image of a privacy label displayed to participants.]

Q8 How confident are you that you know what the label shown above means?
○ Not at all confident   ○ Moderately confident
○ Slightly confident   ○ Very confident
○ Somewhat confident

Q9 I believe the label shown above
○ Strongly decreases the privacy and security risks associated with this specific *[App Name]* app
○ Slightly decreases the privacy and security risks associated with this specific *[App Name]* app
○ Does not have any impact on the privacy and security risks associated with this specific *[App Name]* app
○ Slightly increases the privacy and security risks associated with this specific *[App Name]* app
○ Strongly increases the privacy and security risks associated with this specific *[App Name]* app

Q10 Please explain why you believe the label (decreases/increases/does not have any impact on) the privacy and security risks associated with this specific app
- Please type in your answer.

Q11 Assuming you want to install the *[App Name]* app on your phone, knowing that this app has the label shown above would
○ Strongly decrease your willingness to install this app.
○ Slightly decrease your willingness to install this app.
○ Not have any impact on your willingness to install this app.
○ Slightly increase your willingness to install this app.
○ Strongly increase your willingness to install this app.

Q12 Please explain why knowing that this app has the label (decreases/increases/does not have any impact on) your willingness to install *[App Name]*
- Please type in your answer.

Demographic Questions

D1 What is your gender?
○ Woman   ○ Prefer not to disclose
○ Man   ○ Prefer to self-describe
○ Non-binary

D2 What is your age?
○ 18 – 24   ○ 45 – 54   ○ Prefer not to dis-
○ 25 – 34   ○ 55 – 64   close
○ 35 – 44   ○ 65 or older

D3 Are you a student?
○ Yes   ○ Prefer not to dis-
○ No   close

D4 What is the highest degree or level of school you have completed?
○ No schooling completed
○ Some high school, no diploma
○ High school graduate, diploma, or equivalent
○ Some college credit, no degree
○ Trade / technical / vocational training
○ Associate degree
○ Bachelor's degree
○ Master's degree
○ Professional degree (e. g., J.D., M.D.)
○ Doctorate degree
○ Prefer not to disclose
○ Other (please specify)

D5 Which of the following best describes your educational background or job field?
○ I have an education in, or work in, the field of computer science, computer engineering or IT.
○ I do not have an education in, nor do I work in, the field of computer science, computer engineering or IT.
○ Prefer not to disclose

B Additional Figures and Tables
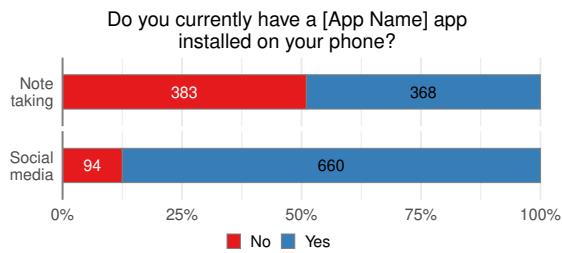
Figure 5: We asked participants if they had an app of this type already installed on their mobile device (**Q3**).
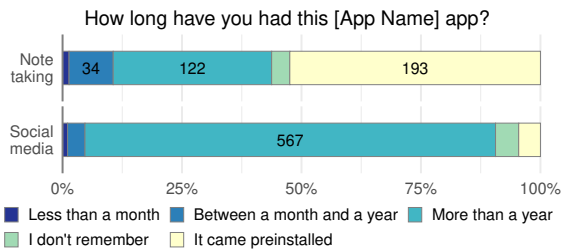
Figure 6: We asked participants how long the app was installed on their mobile device (**Q4**).
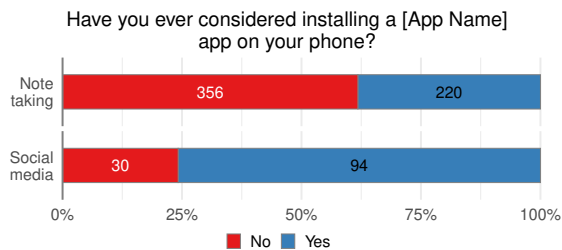
Figure 7: We asked participants if they had ever considered installing an app of this type on their mobile device (**Q6**).
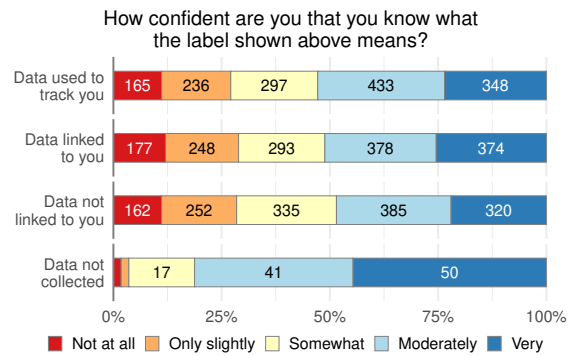
Figure 8: Confidence in label meaning by label privacy type (**Q8**).
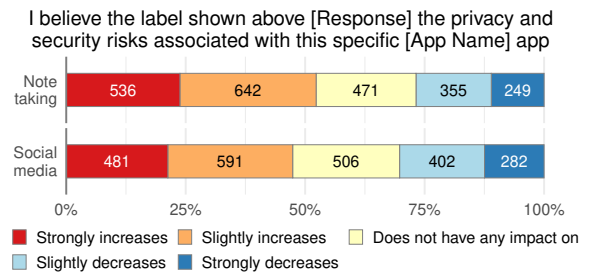
Figure 9: Participant risk perception by application type (**Q9**).
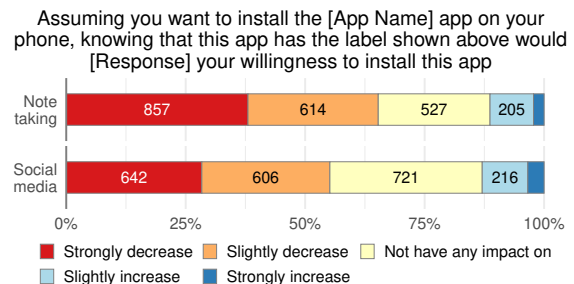
Figure 10: Participant willingness to install by application type (**Q11**).
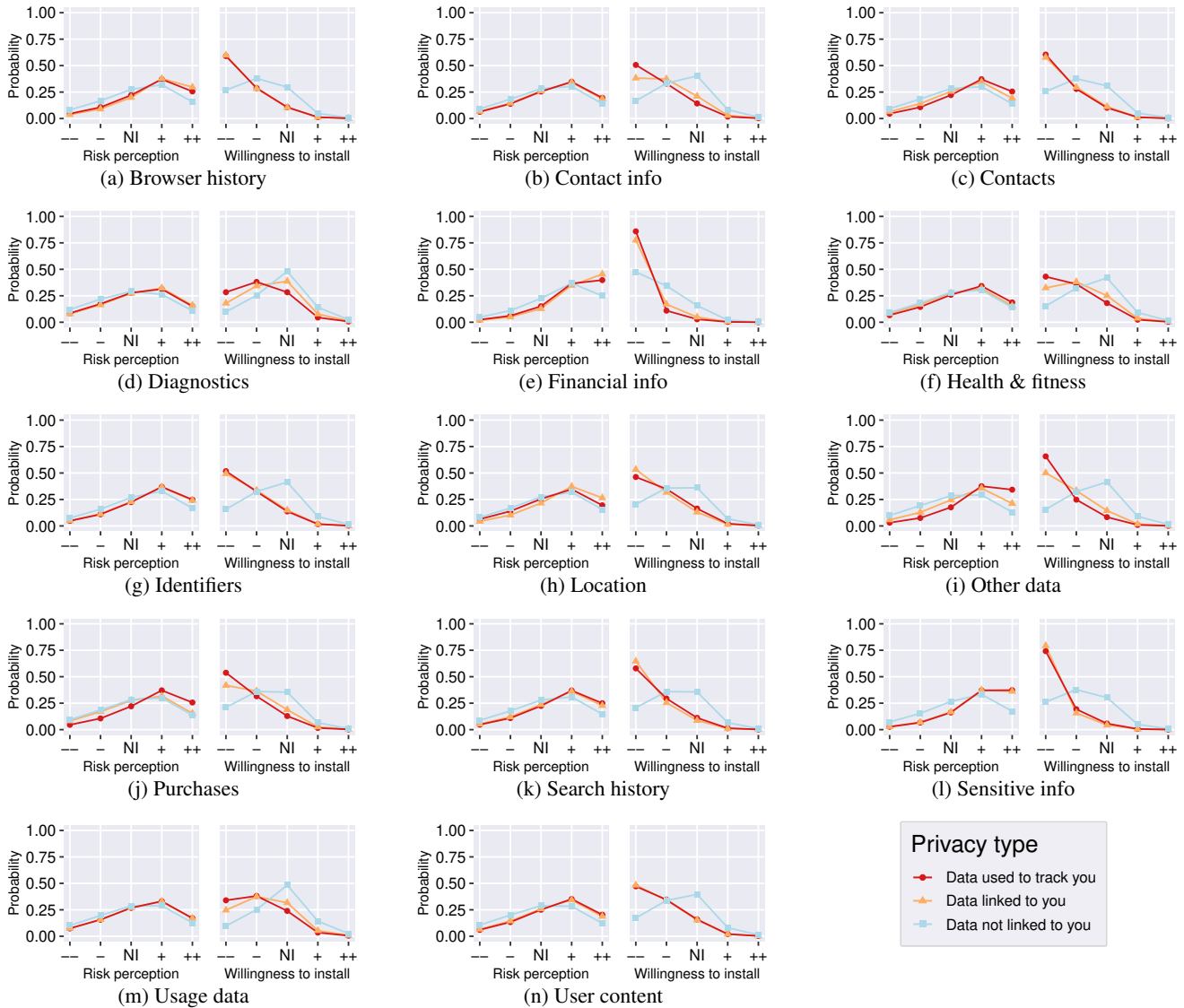
Figure 11: Based on the CLMM parameters, we computed and plotted the probabilities of each data category increasing, decreasing, or having no impact on risk perception (left plot) and willingness to install (right plot). We use the following notation to label the x axes: $--$ is *strongly decrease*, $-$ is *slightly decrease*, NI is *no impact*, $+$ is *slightly increase*, and $++$ is *strongly increase*. For most data categories, the *Slightly increases the privacy and security risks* was the highest probability of the five response categories for risk perception. The exception was the *Financial info* and *Sensitive info* data categories when combined with the *Data used to track you* or *Data linked to you* privacy types, in which case *Strongly increases the privacy and security risks* was the highest probability.
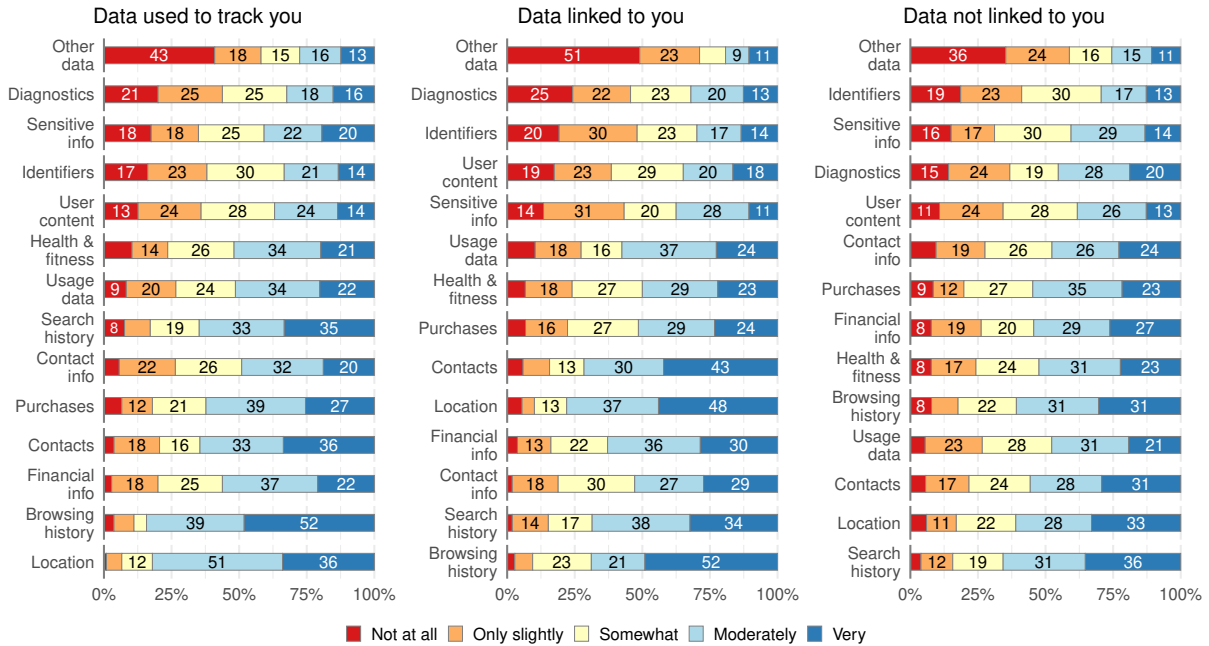
Figure 12: We asked participants how confident they were that they knew what the label shown means (**Q8**).
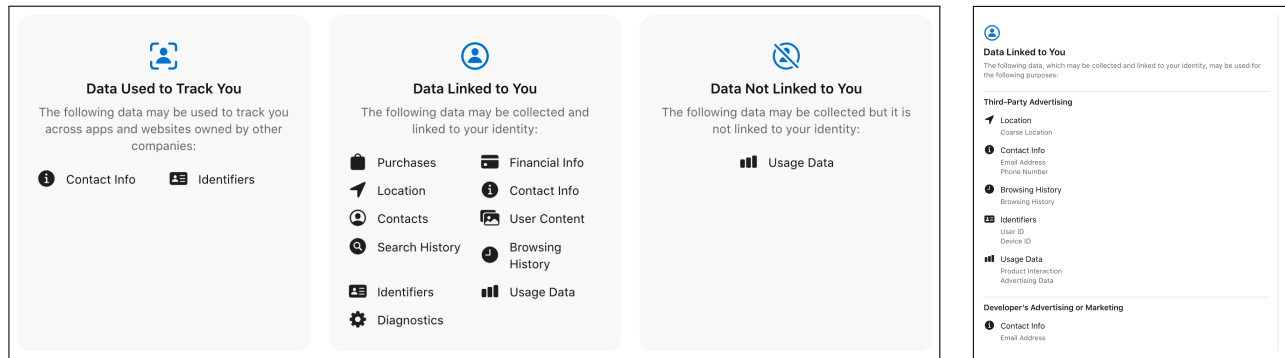
Figure 13: *(left)* An illustrative example of a privacy label from the Apple App Store, and *(right)* an illustrative example of the privacy label details from the Apple App Store. The details display the *Purpose* for the data collection and the detailed information about the *Data Types* collected.

"Say I'm in public...I don't want my nudes to pop up."
User Threat Models for Using Vault Applications

Chris Geeng
New York University

Natalie Chen
Northeastern University

Kieron Ivy Turk
University of Cambridge

Jevan Hutson
University of Washington School of Law

Damon McCoy
New York University

Abstract

Vault apps and hidden albums are tools used to encrypt and hide sensitive photos, videos, and other files. While security researchers have analyzed how technically secure they are, there is little research to understand how and why users use vault apps, and whether these tools meet their needs. To understand user threat models for vault apps, we conducted semi-structured interviews ($N = 18$) with U.S. adult vault app users. We find our participants store intimate media, non-sexual body images, photos of partying and drinking, identification documents, and other sensitive files. Participants primarily used vault apps to prevent accidental content exposure from shoulder surfing or phone sharing, whether in public or with and around close ties. Vault apps were not used to prevent a technically proficient adversary from accessing their files. We find that vault apps prevent context collapse when sharing devices, similar to how privacy settings prevent context collapse on social media. We conclude with recommendations for research aligning with user threat models, and design recommendations for vault apps.

1 Introduction

Vault or secure media storage applications, henceforth referred to as "vault apps", are applications that provide a private media storage repository on a user's phone or mobile device with an additional level of data protection. These applications are commonly used to secure private or sensitive photos, videos, and other documents, protecting them from other users who may have access to regular media storage applications on a phone [1].

Prior work has tangentially noted the use of vault apps for storing sensitive media such as intimate images [20, 34], financial documents, as well as other apps [53], protecting it from other users who may access the owner's phone. Security research has explored traditional threat models to study how secure vault apps are to adversaries with some degree of technical ability [51, 65]. Other work has threat-modeled

vault apps with the user as the adversary, in the scenario that law enforcement is trying to find criminal evidence in vault apps [12, 66]. However, there is a gap in research exploring how existing vault app users make use of the applications and what threats they are actually concerned about.

In this study, we conducted semi-structured interviews (N=18) with adults who use vault apps from a range of backgrounds to identify what motivates people to use these applications, as well as how these applications are used in practice. We asked participants how and why they use vault apps, what features they like or wish the apps had, what threats they are concerned about, and how they found the vault app they use.

Our results include:

1. File assets that are stored on vault apps include intimate media, identification documents, non-sexual body photos, photos of old partners, photos of partying or drinking, medical photos, or conversations.

2. Participants' primary threat models are preventing accidental content exposure from shoulder surfing, when consensually sharing a device, or when a parent is snooping on a phone. They are aware that vault apps may not prevent targeted hacking.

3. While features of usability and security/privacy through authentication can be in tension with one another, participants cite having both features is the appeal of the vault apps they use.

4. Device and photo gallery sharing also produce context collapse given different media has different intended audiences, similar to social media posting [37]; vault apps provide a privacy function similar to granular audience selection in privacy settings, restoring contextual integrity.

By exploring the range of threat models that are considered by these users, we identify ways vault apps protect vulnerable individuals, and we make design recommendations for vault applications to meet user's security, privacy, and usability requirements.

2 Related Works

2.1 Vault Apps and Hidden Folders

Vault apps are digital tools used to protect the privacy of sensitive photos, videos, documents, and sometimes other apps, using encryption, camouflage, and hiding [65]. The average vault app requires successful authentication through a PIN, swipe pattern, or biometrics to access the content stored in the app. Therefore, even if an adversary had access to a person's unlocked phone, they would be promoted to authenticate again to access the files. They sometimes provide decoy functionality by having an app icon that appears as a calculator app on the phone's home screen. There are also social media apps like Snapchat or file apps like iOS Photos that have a secondary feature to password-protect specific files, such as iOS Photos Hidden Album, Google Photos Locked Folder, Snapchat My Eyes Only Album, and iOS Notes locking. When we refer to vault apps, we are also referring to these secondary hidden albums.

2.1.1 Security Analyses of Vault Apps

Security researchers have previously used forensic or security analysis to identify vault apps and extract hidden files from them [22]. Dorai et al. created a tool to automatically extract data from iOS vault apps, to be used by law enforcement [12]; Duncan & Karabiyik, as well as Zhang et al., were able to extract vault app data on Android devices as well [14, 66]. Xie et al. found that for some Android vault apps, they could find the login password or unencrypted information stored in the Android file system [65]. To see if adversaries could extract data without forensic analysis, using the threat assumptions of either unjust search and seizure of civilians by authorities or intimate-partner violence, Ruffin et al. found that "adversaries can infer the existence of most of the popular vault apps and retrieve the stored files with the rudimentary-level knowledge of the Android system" [51]. One limitation of these security analyses was not having empirical evidence on how users actually use vault apps to justify their threat model. Following prior usable security work investigating how different groups threat model their own lives [18, 19, 32, 57, 58], our work seeks to fill this gap.

2.1.2 Vault App Usage

There has been little research on how and why people use vault apps and hidden albums and whether these tools meet their security and privacy needs. Sambasivan & Checkley et al. studied the phone privacy practices of women in South Asia, where the cultural expectation is that they should share their phones with family members so their digital activities may be scrutinized [53]. They found women often employ app locks

(similar to vault apps[1]) to protect social media apps, photos and videos, as well as menstrual period trackers, banking apps, and adult content. Some challenges they had included PINs being discoverable and the presence of an app lock being incriminating. Geeng et al., focusing on adults in the U.S., found that users store intimate photos in vault apps [20].

2.2 Securing Intimate Media

Sharing intimate media, or sexting, has become a common practice: Herbenick et al. found that 27% of adult women and 24% of adult men in the United States sent nude or semi-nude photos of themselves to someone in 2017 [25]. While early literature on intimate messages treated the phenomena as a high-risk, deviant behavior, current research underscores that sexting can be an important part of adult social life that is just as normal as not sexting [15, 30], and can have a positive role in relationship satisfaction [7, 13, 59]. Supporting safety around intimate messages requires acknowledging both "vulnerability and sexual agency" [15], as well as removing patriarchal norms from sexual expression [54], given in general women, sexual minorities, and ethnic minorities bear a disproportionate burden of harms around sexual expression, such as surveillance, harassment, and abuse [8, 9, 52, 55, 56]. And legal scholar Danielle Citron notes that privacy around individuals' intimate lives is a privacy value of the highest order because of its importance to sexual agency, intimacy and equality: "[w]e are free only insofar as we can manage the boundaries around our bodies and intimate activities" [9].

2.2.1 Vault Apps and Intimate Messages

Snapchat is a common app used for sharing intimate messages among young adults [61]. It features a "My Eyes Only" password-protected photo album. Other password-protected storage people have used to store intimate media include Vault and encrypted folders on one's computer [20]. Password protected storage is an often-recommended defense for protecting intimate images [3, 38]. Maas et al. found that, while there has been an incident where high school football players had non-consensually stored nudes of female classmates in vault apps which facilitated posting photos online, non-consensually posting nude images/videos online is not a behavior significantly associated with vault app usage [35].

2.3 Social Media and Device Privacy

Beyond just communication around intimate media, people generally communicate differently based on context and audience [23], and people have different privacy norms in dif-

---

[1] App locks primarily allow phone users to restrict access to any application on their phone by using a password/swipe pattern/biometrics; some vault apps may also have this functionality, but they primarily lock photos and files.

ferent contexts [46]. Marwick & Boyd coined the term "context collapse" to refer to "when social technologies cause a collision of information norms [that] people experience as privacy violations." [37]. This commonly occurs on social media, where a post may be seen by a variety of audiences, not just the poster's intended audience. This can also occur with device sharing. People who trust each other, e.g., family, friends, and romantic partners, often share accounts and devices [39, 47, 48], though device sharing may also happen as obligation [48].

Jacobs et al. found that when a partner in a collocated couples share their device, they may accidentally share private content with their partner. And despite device sharing being common, Wu et al. found that few tools allow intimate partners to maintain both their ideal sharing and security behaviors with devices [64]. Device and credential sharing, compared to healthy relationships, can be adversarial in relationships with intimate partner violence [16, 17, 40]. While device sharing can happen with consent, snooping on smartphones, or non-consensual device access, also occurs: Marques et al. found that 20% of U.S. adults had engaged in phone snooping in a year [36].

Researchers have also explored the tension between parents desiring technology check-ins or surveillance and children wanting privacy from their parents [4, 10, 11, 21, 33, 41, 45, 63]. Hawk et al. found that increased parental privacy invasion led to adolescents telling their parents less about their lives [24]. Cranor et al. found that, while parents believe they should be able to monitor all of their teen's possessions, teens felt their phones should be exempt. In terms of vault apps, various parenting articles caution parents to check their children's phones for vault apps hidden as calculators to find content children are hiding from their parents [2]. To better understand what vault app users, from their perspective, store on these apps, and what their security and privacy concerns are, we conducted our research study which we describe below.

3 Methodology

From July to August 2023, we conducted 18 semi-structured interviews with adults living in the U.S. who used any form of vault app or hidden album folder.

3.1 Recruitment and Participants

To recruit participants, one author posted flyers around a major U.S. city. We also shared flyers with our personal networks, social media sites such as Reddit and Lex, as well as university undergraduate email listservs. Recruitment materials linked to a Qualtrics screening survey, which screened for participants 18 or over and who use vault apps. Given the potentially sensitive nature of discussing vault app storage, other demographic questions besides age were voluntary. Based on responses, we selected participants of varying age, race, income, gender,

ID	Gender	Sexual Orientation	Age in Years	Race
1	man	straight	18-24	Latino
2	non-binary / third gender	gay	25-34	Latino
3	woman	bisexual	18-24	Asian
4	man	gay	25-34	White
5	man	gay	25-34	Asian
6	man	gay	25-34	Black
7	non-binary / third gender	N/A	18-24	Asian
8	woman, non-binary / third gender	queer	25-34	White
9	woman	bisexual	18-24	White
10	man	straight	18-24	White
11	man	N/A	18-24	Asian
12	woman	straight	18-24	Black
13	woman	N/A	18-24	Asian
14	man	straight	25-34	Asian
15	man	straight	18-24	Asian
16	man	straight	25-34	Asian
17	man	straight	18-24	Asian
18	non-binary / third gender	queer	25-34	White

Table 1: Demographic information of interview participants (*N* = 18). N/A means the participant did not specify an answer.

Education	#	Relationship Status	#
bachelor's degree	7	single	7
graduate or professional degree	7	dating	5
some college, but no degree	4	partnered	2
		N/A	3

Household Income	#	Relationship Style	#
$150,000 or more	1	monagamous	16
$100,000 − 149,999$	3	open and monogamous	1
$75,000 − 99,000$	1	N/A	1
$50,000 − 74,999$	2		
$25,000 − 49,999$	3		
less than $25,000	6		
N/A	2		

Table 2: Aggregate demographic information of participants. N/A means the participant did not specify an answer.

and level of education. (Despite our attempts at recruitment, we were not able to recruit any participants over 34.) During the interview debrief, we further asked participants about their sexual orientation, relationship status, and relationship style. Participant demographic information can be found in Table 1, with some demographics presented in aggregate for participant privacy in Table 2.

3.2 Interview Protocol

Two of the authors conducted semi-structured interviews. Some interviews were conducted with both authors present and some interviews were conducted by only one of the authors. We conducted interviews remotely via Zoom for an average of 34 minutes. Participants provided written consent prior to the interviews; the interviewer also provided an overview of the consent form again at the beginning of the call to answer any questions. Zoom calls were recorded with participant consent; we retained audio and deleted video. Participants were compensated with a $30 dollar gift card.

The interview protocol covered what vault apps or hidden album apps the participant uses, how they found out about them, what they store, how they handled the files prior to vault app storage, why do they use them, which features they find useful, what features they wish it had, and any other security or privacy concerns that prompted app usage. If participants used the vault apps to store intimate media, we further asked questions around establishing consent as well as storage duration. The full protocol can be found in Appendix 7.

3.3 Data Analysis

We followed an open coding process. First, we used MacWhisper (an AI-based transcription tool; no third-party person was involved) to transcribe the interview audio. Documents were locally transcribed on the first author's computer. The second author flagged any lines that were unclear and manually corrected them. We anonymized the transcripts.

During the open coding process, two authors double-coded 8 of the same interviews. After double-coding every two interviews, the authors met to discuss code development and resolve disagreements. After the codebook became stable, the authors recoded the first 8 interviews and double-coded the rest, meeting together after every few interviews to discuss and resolve disagreements. Inter-rater reliability (IRR) was not calculated because we double-coded all interviews, and because our research goal is the richness and nuance of different experiences, not counts of how often a code occurred [43].

3.4 Positionality

Three authors identify as queer, and two authors identify as straight. Our paper presents findings, particularly around sexting culture amongst queer people and gay men. While social

science has a history of positioning gay sexual practices as deviant [44], our position on consensual intimate messaging is one of normativity, particularly in gay communities where it can mediate internalized homophobia and loneliness [60].

3.5 Ethical Considerations

Given the potential sensitivity of discussing intimate images as well as other topics, participants were reminded they could skip any question they felt uncomfortable answering or withdraw from the study without penalty. Participants could end the interview at any time and still be compensated. After the interview, participants could request to review their recording and have all or any portion destroyed. No participants requested this. Participant demographic information is presented partially in aggregate to further anonymize participants. This study was approved by [redacted for review] IRB.

4 Results

In this section, we report on participant threat models for using those apps, i.e., what threats they use vault apps for as a defense and what assets they are protecting through vault app storage. We also report on the tool affordances that are important to them or wish to have, how they picked the tool, and what their storage behaviors were before having the tool. Quotes have been lightly edited to remove filler words for clarity.

The vault apps participants brought up included:

1. Hidden albums within photo gallery apps: iPhone Photos, Google Photos. The hidden albums are accessed within each app by traversing the app's albums/library. These hide photos and videos.

2. Snapchat's My Eyes Only album. Photos and videos in this album are stored on Snapchat's servers. If the user forgets the password, the photos are lost forever; there is no recovery process.

3. Samsung's recommended app Secure Folder. It can lock photos and videos, files, and other apps.

4. Third-party vault apps like Secret Photo Album, Photo Vault, Vault, and App Lock.

5. And locked files within existing note apps: Adobe Acrobat, iOS Notes, OneNote.

All of these tools require authentication to access hidden files stored through the app. Their specific affordances are described in more detail in Appendix 7.

4.1 Threats Towards Using Vault Apps

Users installed a vault app to protect their data from shoulder surfing, accidental exposure, and parental device snooping.

Threats Covered by Vault Apps	Threats Not Covered
Shoulder surfing	File-system access
Accidental exposure when sharing device	App data collection
Parental device snooping	

Adversary: Accidental	Adversary: Targeted
Friends	Someone with technical
Family	expertise specifically
People in a public space	after the participant
Co-workers	Vault app corporation
Oneself	

Asset: Sensitive Files
Sexual imagery of others or oneself
Non-sexual body photos
Photos of partying or drinking
Photos of old partners
Sensitive documents, e.g., IDs and password lists
Medical photos or conversations

Table 3: Threats that participants feel vault apps can defend against and cannot defend against.

These threats can be considered with respect to a set of relevant *threat actors*, who have different capabilities that should be considered when designing privacy and security features for the vault apps. A summary of the threats relevant to vault apps are summarised in Table 3. Threats that participants reported as ones not covered by vault apps are discussed in Section 4.3.

4.1.1 Proximity-Based Access

The first threat encompasses people who are in the same vicinity as the user. These may include friends, family, co-workers, or strangers in the same space. These users do not have any control over the device itself; however, they are likely to be able to see the content of the users' screen through "shoulder surfing" or because the user is sharing the screen. To prevent this, vault apps move media from applications which would regularly include them (e.g. camera roll, file viewers) into other storage.

> When I'm in public, I'd rather not have some of those [intimate] images that are far back to be shown as I'm scrolling, 'cause I take a lot of public transit. I'm usually around a lot of people. So I wanna make sure that those images aren't accidentally shown when I'm in some of those more social situations. —*P2*

> If [I] open up Snapchat, say I'm in public. I'm

recording a drag show or something. I don't want my nudes to pop up. So I just don't want people to see [these photos] when I'm scrolling through. —*P18*

4.1.2 Physical Access to Device

The second threat includes people who have physical access to the user's device. Some of our participants would share their phone with a friend or family member to share some memories, who may accidentally scroll past the intended shared photos.

> Sometimes I share my camera roll with other people. So I wanna make sure that some of those images aren't just shown. —*P2*

P3 and P17 were also concerned about parents accidentally seeing content while they intentionally snooped through one's phone. And some participants had photo memories of a time they would not want to be reminded of unless they were in a specific mood.

Vault apps provide a secondary locking mechanism, which is often (but not always) distinct from the devices' authentication mechanism. This protects against this threat by adding an additional layer of authentication, that users with physical access cannot bypass by accessing the unlocked device alone. Vault apps also separate sensitive content from general folder content, making it more difficult to accidentally send the wrong photo. P9 stated, "I don't want it in my Apple photos because I could accidentally send this to someone."

4.2 Assets Protected by Vault Apps

In general, participants stored types of sensitive files that, when shown to an unintended audience, could bring up feelings of embarrassment or shame, loss of dignity, questions or lectures from family, or old memories in oneself. For example, P17 did not want certain friends, who he describes as "something common here and around in India...they only talk with boys, they only interact with boys, and they have some sort of exclusion from women," to see photos of him with another girl because:

> When they find out those pictures, they might make a fuss around the classroom....They would shout around, shout the girl's name when I'm around, shout my name with the girls around....It's just a friendship thing, but they take it as a crush or whatever....So yeah, [I use a vault app] to avoid that annoying thing and not to embarrass that person as well.'

We discuss specific stored content below.

4.2.1 Asset: Sexual imagery of others or oneself

Many participants mentioned using vault apps or hidden folders to store not-safe-for-work (NSFW) sexual photos of themselves or others. The common threat they were concerned about was accidentally revealing those photos in a public place, in a professional setting, to family or friends, or anyone other than the participant themselves and the person they intend to share it with. These could be nude or semi-nude photos, as well as fetish-related photos. People mentioned using Snapchat's My Eyes Only and Hidden Photos Album in particular.

P6 and P8 in particular mentioned potential harms to others accidentally exposed to one's intimate imagery could cause. P6, a gay man, talked about not wanting his women friends to see his photos:

> But the idea of a woman seeing unsolicited dick pics [is] then just sort of me perpetuating an already existing system or [...] harassment.

And P8, a queer non-binary person said,

> In the work context it would be really bad and I think in any context it could be really alienating for someone, and kind of sexual harassment.

Some participants, while they used vault apps to protect against accidental exposure, were not as concerned if it did occur. P4, a gay man, stated, "I'm also just like very sexually liberated person. So I don't really like, think of too many bad outcomes." He said that while he has the concern of a targeted adversary getting a hold of his images, or his boss accidentally seeing his images, he finds these scenarios quite unlikely.

Storage Practices. Some participants discussed storing a partner's intimate photos prior to saving, while others who sext in casual relationships said that being sent an intimate photo implies consent to store it.

> People, if they don't want something to be saved, would send it over like Snapchat or like the Instagram disappearing photo messages. —*P4, gay man*

P9, a bisexual woman, talked about desiring to discuss storage boundaries:

> I'm bisexual and I think with women, it's a little bit more of an open conversation of what is expected, 'cause you're both a little bit more aware and a little more scared. I think with men, when I talk with men, they don't really care as much what happens with their photos as I do with me. So it is more like on my end, like, hey, don't screenshot, don't do this, don't do that. And obviously I'm like, if you would like, if you want me to do the same, I will, but most of the time they don't care in the sense that women tend to care about their photos.

Some participants did not think about the storage duration of intimate images, while P1, P4, and P9 said they deleted intimate images of partners at the end of the relationship.

4.2.2 Asset: Non-sexual body photos

Several participants mentioned storing non-sexual photos of their bodies for various reasons. This included gym body progress photos, gender-transitioning photos, and swimsuit photos. P3, a 18-24 year old woman, talked about putting body photos from when she had an eating disorder into a vault app so she does not accidentally see it:

> That's another reason that I don't really look back on My Eyes Only 'cause, at least where I'm at right now, I'm trying not to be in that mindset. But at the same time, I don't know that I would necessarily delete it because sometimes it is helpful to look back and [see], this is what I was thinking.

P7 took photos of themselves to track their bodily changes as they started hormone replacement therapy. They keep photos in the iOS Hidden Album to prevent an unintended audience from seeing it, which would cause:

> Embarrassment or loss of dignity or decency, I think particularly because it's my body going through transition, it's something personal to me. Or it's not a version of myself that I present to people now.

They also do not want to accidentally see those photos because as they mentioned, "I don't really want to look at my body", so using a vault app requires conscious access to see the photos.

P8 also put SFW selfies in their Hidden Album:

> I think I felt like I had too many and they didn't need to be in my regular camera roll....But they were kind of clogging my camera roll.

4.2.3 Asset: Photos of partying or drinking

Some participants stored photos of drinking or going to parties with friends that they did not want their more conservative families to see.

P1, an 18-24 year old man, said:

> The earliest memories in [Snapchat My Eyes Only] are from the stupid high school parties where I didn't want my aunt or uncle [to see], if I'm throwing in a picture of a memory or whatever to see [me] drinking.

P16 and P17 mentioned not wanting parents to see photos of alcohol. P17, an 18-24 year old man, talked about using Secure Folder because while he does not drink, he keeps photos of hanging out with friends who drink, which his parents would ground or lecture him over.

I had to hide [the photos] from my parents because they find [alcohol] unfavorable in their eyes. So me coming from a strict household, I have regular checkups on my phone, even though I'm [18-24] now, I have regular checkups on my mobile phone. My parents do that regularly.

4.2.4 Asset: Old Partners

Several participants mentioned not wanting a new partner to see photos of old partners. P6 said,

It's not like I'm fearing that relationship would somehow be torn asunder because of the presence of those photos, but it does create a tiny little thorn of just sort of like, ooh.

Some participants kept photos of exes to look back on as an old memory. P5 described:

I don't feel like if I break up with someone I need to remove that part from me. And I always talk with my current partner about [my] dating history. For example, for some photos, I would just hide them for the purpose of keeping the memory or separating them from the major album.

4.2.5 Assets: Sensitive Documents

Participants mentioned using vault apps for important documents, such as bank statements, identification documents (ID proofs) and for P16 from India, "certificates to prove our caste and religion."

P11 started using a vault app because he was looking for a safe and easily accessible storage mechanism for IDs due to this incident:

In 2020 when I was traveling, in [the] airport we have to show our ID [as] proof that our names and the passport, or any ID proof, matches. I had an e-ticket and I had my ID proof as a PDF in my phone. I was unable to find [the] PDF because I had saved it in my local chats and everything in my WeChat. There were people in the line and everything so it was a very hassle moment for me.

The vault app made it easy for him to find his IDs on his phone. And his concern if friends or family got access to his IDs was that information would get to a loan shark who would call and harass him, given that happened to a friend.

P1 also uses Dropbox's locked files to store sensitive financial information regarding a project's donors. P1 uses iOS Notes, P13 uses OneNote, and P16 uses AppLock for storing a locked note filled with various passwords, functioning somewhat similarly to a password manager.

4.2.6 Asset: Medical photos or conversations

P2 mentioned storing photos of a skin problem when they were having intestinal issues in their My Eyes Only Album, because they considered it not-safe-for-work:

And then when I would go to the doctors, if they needed to see a visual from previous irritation I could also share that with them.

P9 talked about using the Locked Folder in Google Photos in high school for saving screenshots of conversations to discuss with her therapist. She started using it because her friends would play Photo Roulette, which would access her Camera Roll. She moved photos she did not want her friends to see through the game to the Locked Folder, including photos related to her bipolar disorder diagnosis:

I don't need people learning things about me that I wouldn't tell them willingly because they see something that [..] I didn't want them to see. So I did it for myself. Some things were moved there once I got word [of] things from other people, like you wouldn't want to know that, like when I had a mental health diagnosis. So I was hiding things that were related to that because I didn't want people to know [...] I have bipolar disorder. So other people were like, 'That's not a common enough one. You should hide anything related to that [...] that's a shameful diagnosis or disorder to have'.

She stopped using the tool in college since her parents had stopped going through her phone then, and also,

I didn't feel my privacy was being invaded as much. I felt more comfortable setting boundaries. So I didn't think I needed to hide everything. And I think also I was like, okay, if I stopped hiding things on my phone, it'll make me feel more secure in telling other people things that maybe, because I was able to hide things that I wasn't able to do beforehand.

4.3 Threats Not Covered By Vault Apps

Participants had other security and privacy concerns for their assets that they did not use vault apps to cover.

4.3.1 Targeted Adversary

Participants also sometimes had concerns of a targeted, skilled adversary, but were generally aware this app would not protect against that kind of threat. In terms of a targeted hacker, P9 stated "But I'm like, so many people use Google. What is the odds that it's mine that gets hacked?"

P6 said,

I really don't fear too much of somebody going through my phone. What person would be on that chaotic mission of "I'm gonna go try and find all of your [nude] photos right now". If another guy gets in, frankly, I don't have that many straight male friends. If he does get in, I'm just like, well, that's on you. I don't know how you find yourself there, but that's what you got. I may be delusional, but I don't really really fear bad scenarios. I stopped using it once I was a sophomore in college, because by that point [..] I didn't feel my privacy was being invaded as much.

P8, when asked why someone accidentally seeing a photo rather was the bigger concern over a targeted hacker, said,

> I think both should be concerns and I should probably be more concerned about my stuff getting hacked, rather than me accidentally showing something to someone. But I think that I'm a little bit lazy and I tend to not prioritize real privacy concerns I should have. Just for the sake of convenience I skip over and I don't make the effort to really protect things.

4.3.2 Company/Government Data Collection

Participants were concerned about the company of the tool they were using collecting data from them as well. P7, P9, P10, and P13 stated they were concerned about Snapchat having access to their sensitive photos and not knowing how they are using the photos or how well they are protecting those photos.

For P5, he preferred using the iOS hidden album because they did not want to download a third-party vault app and have another company collecting information from him. For P14, he trusts Apple well enough to use their tools, but notes, as an international student, because of the Patriot Act:

> [The government] can basically, no matter the level of encryption or anything that you have, no matter if you're using a phone of Apple, Samsung or anything, they can literally get into your phone.

P9 stated about Google Photos Locked Folder,

> With [the app], I think it was just more concerning 'cause that was at the start of when people started to learn about how data was being mined from us and being sold to companies and all that stuff, that was [..] when people started freaking out about that.

However, she concluded that she already uses many Google products and they already have a lot of information on her.

4.4 Vault App Affordances and Features

We discuss the affordances that met participant needs, and affordances they wished were included. A summary of the af-

fordances that participants mentioned positively can be found in Table 4. Table 5 and Table 6, which lay out which apps participants mentioned have which affordances, can be found in Appendix 7.

4.4.1 Authentication

Customizable Password. P17 mentioned being able to use a different password than his phone password, which makes him feel safer because,

> Since my parents do regular checkups, they know my password to open the phone, but for the secure folder that I have hidden the password from them.

P13 was concerned about not having this feature with Hidden Folders, because knowing how to unlock the iOS phone would allow unlocking the Hidden Folder. For P1, the customizable password determines what folder is opened.

> Based on whatever pin you enter, it'll direct you to a different folder without revealing that there are any other keys in the name system.

Biometric Authentication. Some participants like the ease of biometric authentication, whether through fingerprint or FaceID. P11, discussing opening identification documents said,

> I use frequently the fingerprint one. Because if I have multiple bags in my hand at the airport, keeping them down and then typing the password is going to get pretty complicated.

But P1, P8, and P16 were concerned that biometrics makes authentication too easy for a potential adversary to unlock their files, which we discuss more in Section 4.4.4. Because Google Photos Locked Folder uses the default authentication for the phone, P16 wanted to be able to change the default biometric authentication to a PIN.

4.4.2 Discreteness

P3, P5, P7, P9, and P16 mentioned liking their tool of choice because it was difficult to find in a phone. P9 said they like Snapchat's My Eyes Only because, "[My parents are] not going to understand how to access it." For P16, they liked AppLock because the icon on the phone looks like a calculator. P17 wished that Secure Folder would disguise itself as another kind of app, such as a calculator.

P3 described wanting something similar specifically for locking notes or text messages, in a way "that's like discreet enough to not be a clear sign of if you're hiding something" so that it does not alert her parents or friends that she is hiding something. P3 did not want to download a separate vault app because of questions its existence could invite:

Positively-Mentioned Affordances
Customizable password different from phone authentication
Biometric authentication (fingerprint or FaceID)
Different folders opened by different customizable password
Disguising app icon (e.g., Calculator)
Hidden album is within an already used app (e.g., iOS Photos, Snapchat)
Vault app is separate from an already used app
Locked photos are hidden from Memories (e.g., Snapchat)
Built-in sharing functionality to other social media or messaging apps
Folder creation

Table 4: Summary of existing vault app affordances that participants positively mentioned. Some affordances are contradictory: some participants liked the ease of access that biometric authentication provided, while other participants did not feel as secure that the vault app could only use the device's default authentication, which was often biometrics. In addition, some participants preferred using a hidden album within an existing photo or social media app because it was convenient or because friends/family would not find it suspicious, while other participants preferred having a separate vault app to not create folder artifacts in commonly used apps.

> If my parents or my friends were to see, oh, you have this app just for hiding photos or certain things on your phone, I think that would have been maybe a little weird, felt like I was being overly secretive.

While Grindr does not have a locked photo album feature P4 did like that its existing Albums do not show a preview of the photos contained inside, so he can discretely share albums while other people are around without exposing any photos. P2 and P13 liked how Snapchat, for photos placed in the My Eyes Only album, are automatically removed from Snapchat's Memories, which is where photos are generally saved.

4.4.3 File Separation

P6 describes managing his intimate images as,

> What I want is just take all these photos and pull them out of the general pool. But the real value is them just not being [in] the general pool.

This file separation is enough for folks' threat models, but as others have pointed out in addition to P8, they don't consider it "secure".

> I think from my understanding is it's just putting it in a different album on my phone that is not with the rest of my camera roll, but it's not really protecting it in any way.

For P1 and P14, who use Secret Photo Album to store intimate images and Vault to store non-intimate images of former partners respectively, they like using apps that are not a default on their phones because it does not produce artifacts in popularly known locations. P1 stated he feels Secret Photo Album is more secure than using iPhone Hidden Album because his friends would know of the latter but not

the former. And P11 says he likes using a separate vault app instead of Android's built-in locked Gallery album because,

> So for [IDs or other photos], I started using it in Secure Folder[...]it's also protected from [the] normal Gallery....I didn't want to get into hassle of those because if someone wanted to access [an album in] the Gallery, lots of questions can pop up.

4.4.4 Ease of Access

P11 prefers PhotoVault to Secure Folder for storing IDs because Secure Folder needs a second click into a folder to open up images, while PhotoVault shows images upon successful authentication.

Several participants began to or liked using their chosen tool because it already existed on their phones.

> It's just convenient that it's already on my phone. I didn't have to download anything else. —P7

> It's convenient in the sense that I have Snapchat on my phone. I use Snapchat pretty often, so it's easy access. —P9

Some participants saw this ease of access as a security cost. P3 wished that My Eyes Only would have an additional pop-up question of "Are you sure you want to send this?". P8 wished there was an additional step to authentication for the iOS Hidden Album after FaceID.

> I guess I have one concern if I was asleep or something and I don't know if face ID works when your eyes are closed. Or if I accidentally clicked the hidden album and then it immediately scanned my face and opened it up. It would be nice to have extra questions or require extra actions.

P1 felt similarly about accidentally unlocking the iOS Hidden Album with his finger and wanted an extra prompt asking for authentication rather than automatically doing so.

4.4.5 Ease of Sharing

While P2 liked how easy it was to share within Snapchat and download photos from the My Eyes Only album, a common desire amongst participants who shared intimate messages was being able to easily share a photo from a different vault app or hidden album to a messaging or social media app. P4, P5, and P6 wished more apps would have access to the iOS Hidden Album when uploading photos. P14, who has Vault on his iPhone, wished there was an easy way to locally transfer files from it to his PC without having to use Vault's Cloud feature.

4.4.6 (Absence of) Backups

P1 was concerned about losing his stored photos if he forgot his My Eyes Only PIN. He could reset it, but Snapchat will delete the images in the album.

Conversely, P3 was concerned with not wanting Hidden Album photos to be backed up with iOS. P14 complained that by default with iCloud featured turned on, the Hidden Album photos will also get stored, which she does not want and has turned off. They would not want that automatically.

P7 stated he wanted more transparency around what happens to files with iOS backups:

> I think, in particular, like when you back up your phone, I don't know where those photos go like on your computer....But I guess like [I want] more transparency and like, if they're committed at all to protecting what you have in there, because I guess the assumption for me as a user that like hiding photos would have, would that be like things are private or people like want them to be hidden for some reason or another that I would presume be private.'

4.4.7 Desiring more "Security"

Meanwhile, P11 had a specific request for more security. If someone opened and failed to authenticate for Photo Vault, he wanted it to discretely take a photo of the person so he could find it later. P14 talked about an app feature that would automatically filter more private photos from the general photo pool based on a country's specific norms of what is considered private. And P17 wanted an explanation of My Eyes Only when he first found it:

> If they would just give me a little tutorial or something, break down of the feature and its security, that would probably be good.

4.4.8 Other Desired Features

P2 and P14 wanted more editing features. P2 wanted to be able to send blurred out photos through My Eyes Only, similar to how iMessage can send blurred photos where the blur is slowly removed. P1 wished Secret Photo Album could store files other than just photos or videos, such as PDFs of important documents.

P11 and P13 did not like how the free version of vault apps have a lot of ads. While none of the participants mentioned paying for a vault app, we note certain affordances only being available in paid app versions in Table 5.

4.5 Tool Discovery Process

Participants mentioned finding the tool they use either through friends/family, personal discovery while going through an existing app, or doing an Internet search for a vault app (similar to how people have looked for security advice [50]).

P1, P5, P7, P8, P13, and P17 found the Hidden Album feature while using the built-in Photos app on their iPhones. P2 and P12 said they found the My Eyes Only album through their regular use of Snapchat, while P10 said he always knew that tool was there. P14 discovered Adobe Acrobat's locked file feature when he was sent bank statements with a password. P15 had Secure Folder recommended by their Samsung phone, and P16 found Google Photos' locked folder through their regular use of the app.

P1, P3, P9, P11, P13, P14, and P15 mentioned doing an Internet search to either find a tool to meet their needs or to see if an existing tool they had (like OneNote or iPhone Photos) could hide or protect certain files. P11 compared reviews of different vault apps to make a download decision. P14 and P16 mentioned discovering the vault app they use by seeing their friends use a vault app and asking about it.

4.6 Pre-Vault App Behavior

Participants had different asset storage practices before using a vault app or hidden album. P1 would hold his phone close while scrolling past sensitive partying videos rather than scrolling with his phone screen out in the open. P3 would either not take "embarrassing" photos of herself or delete them right after taking them. P1, P8, and P17 did not take nude photos before having a vault app.

P16 would delete photos of himself with friends drinking after taking a look at the photo. P4, P5, and P9 would delete their nudes from their phone after sending it to someone. P9 stated,

> More so with nudes...it would be taken directly on Snapchat and immediately deleted. And then I realized how annoying that was, but I didn't want them so openly on my camera roll....Then once I

discovered [Snapchat] For My Eyes Only existed, I was like, okay, this is great.

P8 would not have stored any or would have deleted nudes of themselves and others. For other participants, before using a vault app they had their photos or documents stored on other apps, including their regular Camera Roll, WeChat files, Gmail files, Grindr, and Snapchat.

5 Discussion

5.1 Perspective Determines Technological Security

While prior work has explored vault app security either from the threat modeling standpoint that the vault app user is the adversary [12, 14, 66], or from the threat modeling standpoint that a vault app user faces law enforcement [22] or IPV [51], our work begins from learning user threat models towards using vault apps. We show how our participant pool uses vault apps towards less capable adversaries than a targeted hacker or law enforcement: our participants were primarily concerned with stopping unintended exposures to the public or close ties around them (which prior research also recommends as the use case for vault apps [22]). While they had an understanding that vault apps would not stop a targeted adversary, for preventing accidental exposure they found vault apps to meet their needs (though any technology can never be proven to be fully secure [27]). Herley writes, "In the absence of actual compromise data the security community often speaks of worst-case risk" [26]. Prior vault app security analyses focused on worst-case risk, but our findings show that for users with different adversaries in mind, the security of vault apps needs to be evaluated differently. For research, evaluating tool security for what threats users care about is just as important as evaluating security for worst-case threats. Therefore, it is important to conduct user studies so users can define their own security concerns and contexts for what mitigations are useful to them, beyond just security experts determining threats for users, as prior work has also shown [19, 58].

5.2 Context Collapse and Device Sharing

While context collapse was initially coined to refer to posts on social media where multiple audiences may see content [37], our results show how context collapse also exists amongst device sharing and screen sharing. Our participants have opened their devices and photo galleries in professional settings, amongst friends and family, and in public, which are not always the intended audiences for all device content. For example, while one may have consensually stored intimate images on one's phone, it would be inappropriate to be seen in a workplace. Given that photo galleries and file apps store media for different purposes and audiences, from public to private, personal to professional, similar to social media, tools are needed to support device privacy similar to how privacy settings prevent context collapse on social media.

Jacobs et al. describe accidental content sharing when partners have access to each other's phones as breaking contextual integrity [28], or breaking adequate privacy based on norms in a specific context [46]. Vault apps allow people to regain contextual integrity through more granular app-level authentication, providing selective content visibility. As shown in Section 4.6, some participants would not have stored intimate media or party memories with friends before using a vault app. This shows how vault apps can provide the necessary granular device privacy that Wu et al. states is lacking on devices for privacy-supporting consensual device sharing [64].

5.3 Vault Apps For Vulnerabilities

While contextual integrity can explain the desire to keep certain media private based on social norms [28, 46], McDonald & Forte argue that privacy theory should move away from protecting norms towards protecting vulnerable populations, who are "not only more likely to be susceptible to privacy violations but whose safety and wellbeing are disproportionately affected by such violations" [42].

We had several participants who had privacy vulnerabilities, either on a societal, community, or familial level: P17 wanted to keep photos of drinking private from his conservative Christian family, and P9 had expectations of keeping her bipolar disorder diagnosis quiet in her town, while also having concerns about her intimate media being exposed as a woman. In the former's case, a parent's desire to know everything about their child comes into tension with their child's desires and values; as Levy & Schneier note, "The balance between essential caretaking and privacy invasion can be unclear" [33] . Regardless, in a parent-child relationship the child often has less power than a parent, even as an adult. And in P9's case, there is much research supporting the higher degree of harm women face from their intimate photos being exposed [5, 6, 8, 31]. Vault apps can provide privacy protection in these contexts of differential power. While vault apps can be used for harmful purposes [12, 35], and some parental articles have flagged it as something to look out for on a child's phone [2], its existence on a phone should not necessarily imply harmful behavior on the user's part because it is also a tool that supports privacy for vulnerable populations.

5.4 Design Recommendations

Most of our participants wanted both security and privacy, as well as usability with their vault apps. Sometimes these values came into tension with one another, e.g., biometric authentication provides easy access but introduces the concern of easy access by an adversary. In accordance with these tensions, we make some recommendations for vault app design. Vault apps

and hidden albums should allow authentication choices in addition to or instead of the phone's default authentication. This includes both PIN and biometric options, as well as options for friction pop-ups to prevent accidental access to the app from a fingerprint or FaceID.

We also note that while all the apps mentioned were free, some had ads while others had paywalled features. Both Android and iOS have default hidden/locked albums within default photo apps (iOS Photos and Android Gallery). If one wants other functionalities like decoy vaults, one has to turn to another app. Prior work has discussed the digital divide between higher-income and lower-income users, with the former having better access to paywalled privacy or security services [49]. While the default vault app feature of authentication-protected storage is available on iOS and Android phones, other features require payment. We recommend phone OS developers continue developing vault app features for default apps to support users with less financial privilege.

Finally, several participants mentioned not fully understanding how the technical aspects of vault apps work, in terms of security. Vault apps could provide an overview of encryption and other security features in-app, similar to the Privacy Center on Facebook. Some participants also mentioned difficulty in determining whether hidden media is backed up on a cloud server or another device. Some vault apps like Snapchat My Eyes Only do not have a recovery process for photos if the password is lost; photos are deleted if the password is changed. Meanwhile, hidden photos on the iPhone are automatically backed up to iCloud if Photos is enabled for iCloud sync. To better provide information on data storage and access, vault apps could have a privacy nutrition label [29] that notes if there are any other devices or servers the files are stored on. Also, for people who have a photo or file backup sync turned on, there should be granular settings to bulk-remove an album/folder from being backed up.

6   Limitations and Future Work

We do not perform a technical analysis of vault app security, as that is out of the scope of this paper. As our research method is qualitative, we cannot provide results on the frequency of behaviors, threat models, or app usage; instead, our interview results provide rich and contextualized insights on participant privacy concerns, motivations, and how that affects their behaviors. Future work should take quantitative approaches to understand the frequency of vault app usage and storage of different media types. Moreover, future work should study how vault apps represent their security to users and whether this maps to actual technical practice.

While we recruited a diverse U.S. population across gender, sexual orientation, race, education, and household income, our participants were almost all monogamous, and participant ages ranged from 18-34. Future work is needed to determine if this reflects the typical vault app user age range. Our par-

ticipants being largely LGBTQ+ is likely due to additional recruiting through queer communities; other groups that we did not specifically recruit from may have different usage patterns with these tools. Also, we cannot speak to mental models of people who have considered using vault apps for shoulder surfing and accidental exposure threats but decided not to use them. We also did not specifically recruit for certain vulnerable populations, such as IPV survivors, who may have different or differently prioritized security and privacy requirements [62], e.g. stealthiness or deniability. Future work should study how different vulnerable populations may use vault apps to understand their more specific threat models.

Finally, we only describe vault app usability and affordances for the tools our participants mentioned. Future work should explore user evaluations of other vault apps and other features they provide, such as Face Down Lock, which closes the vault app and opens a different app when the phone is placed face down.

7   Conclusion

To understand user threat models for using vault apps, we conducted semi-structured interviews with 18 adults in the U.S. who use vault apps or hidden folders. We found the primary threats participants use vault apps to defend against are accidental content exposure through shoulder surfing, when consensually sharing a device, or when a parent is snooping on a phone. Participants stored files including intimate media, identification documents, non-sexual body photos, photos of old partners, photos of partying or drinking, and medical photos or conversations. Given phones store a range of public to private, professional to personal media, we show how vault apps can prevent context collapse and protect contextual integrity when sharing devices. We also show how vault apps can preserve privacy for vulnerable individuals and its existence should not by default imply harmful behavior. We conclude with design recommendations to improve balancing the usability/security tension of vault apps.

Acknowledgments

References

[1] Vault - hide pics, app lock - apps on google play. `https://play.google.com/store/apps/details?id=com.netqin.ps&hl=en_US&gl=US`. (Accessed on 02/14/2024).

[2] Parents warned of vault apps on their children's smartphones. https://www.moms.com/fake-calcula tor-app-kids-hide-photos/, 2021. (Accessed on 02/09/2024).

[3] Toward safer intimate futures: Recommendations for tech platforms to reduce image based sexual abuse - european sex workers' rights alliance. https://www.eswalliance.org/toward_safer_intimate_futures_recommendations_tech_platforms_reduce_image_based_abuse, 2023. (Accessed on 02/07/2024).

[4] Mamtaj Akter, Amy J Godfrey, Jess Kropczynski, Heather R Lipford, and Pamela J Wisniewski. From parental control to joint family oversight: Can parents and teens manage mobile online safety and privacy as equals? *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–28, 2022.

[5] Rikke Amundsen. 'The Price of Admission': On Notions of Risk and Responsibility in Women's Sexting Practices. In Karen Lumsden and Emily Harmer, editors, *Online Othering: Exploring Digital Violence and Discrimination on the Web*, Palgrave Studies in Cybercrime and Cybersecurity. Palgrave Macmillan, 2019.

[6] Rikke Amundsen. The turn to trust: adult women, hetero-sexting, and the use of trust as sexting risk mitigation. *Feminist Media Studies*, pages 1–16, 2023.

[7] Melissa Burkett. Sex (t) talk: A qualitative analysis of young adults' negotiations of the pleasures and perils of sexting. *Sexuality & Culture*, 19(4):835–863, 2015.

[8] Danielle Keats Citron. *Hate crimes in cyberspace*. Harvard University Press, 2014.

[9] Danielle Keats Citron. Sexual privacy. *128 Yale Law Journal 1870 (2019); U of Maryland Legal Studies Research Paper No. 2018-25*, 2019.

[10] Lorrie Faith Cranor, Adam L Durity, Abigail Marsh, and Blase Ur. {Parents'} and {Teens'} perspectives on privacy in a {Technology-Filled} world. In *10th Symposium On Usable Privacy and Security (SOUPS 2014)*, pages 19–35, 2014.

[11] Alexei Czeskis, Ivayla Dermendjieva, Hussein Yapit, Alan Borning, Batya Friedman, Brian Gill, and Tadayoshi Kohno. Parenting from the pocket: Value tensions and technical directions for secure and private parent-teen mobile safety. In *Proceedings of the sixth symposium on usable privacy and security*, pages 1–15, 2010.

[12] Gokila Dorai, Sudhir Aggarwal, Neet Patel, and Charisa Powell. Vide-vault app identification and extraction system for ios devices. *Forensic Science International: Digital Investigation*, 33:301007, 2020.

[13] Michelle Drouin, Manda Coupe, and Jeff R. Temple. Is sexting good for your relationship? It depends. . . . *Computers in Human Behavior*, 75:749–756, 2017.

[14] Michaila Duncan and Umit Karabiyik. Detection and recovery of anti-forensic (vault) applications on android devices. 2018.

[15] Nicola Döring. Consensual sexting among adolescents: Risk prevention through abstinence education or safer sexting? *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 8, 01 2014.

[16] Diana Freed, Jackeline Palmer, Diana Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. "A stalker's paradise": How intimate partner abusers exploit technology. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, New York, NY, USA, 2018. Association for Computing Machinery.

[17] Diana Freed, Jackeline Palmer, Diana Elizabeth Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. Digital technologies and intimate partner violence: A qualitative analysis with multiple stakeholders. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–22, 2017.

[18] Alisa Frik, Leysan Nurgalieva, Julia Bernd, Joyce Lee, Florian Schaub, and Serge Egelman. Privacy and security threat models and mitigation strategies of older adults. In *Fifteenth symposium on usable privacy and security (SOUPS 2019)*, pages 21–40, 2019.

[19] Christine Geeng, Mike Harris, Elissa Redmiles, and Franziska Roesner. "Like lesbians walking the perimeter": Experiences of US. LGBTQ+ folks with online security, safety, and privacy advice. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 305–322, 2022.

[20] Christine Geeng, Jevan Hutson, and Franziska Roesner. Usable sexurity: Studying People's concerns and strategies when sexting. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*, pages 127–144, 2020.

[21] Arup Kumar Ghosh, Karla Badillo-Urquiola, Shion Guha, Joseph J LaViola Jr, and Pamela J Wisniewski. Safety vs. surveillance: what children have to say about mobile apps for parental control. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2018.

[22] Alissa Gilbert and Kathryn C Seigfried-Spellar. Forensic discoverability of ios vault applications. *Journal of Digital Forensics, Security and Law*, 17(1):1, 2022.

[23] Erving Goffman. The presentation of self in everyday life. In *Social Theory Re-Wired*, pages 482–493. Routledge, 2016.

[24] Skyler T Hawk, Loes Keijsers, Tom Frijns, William W Hale III, Susan Branje, and Wim Meeus. "i still haven't found what i'm looking for": Parental privacy invasion predicts reduced parental knowledge. *Developmental Psychology*, 49(7):1286, 2013.

[25] Debby Herbenick, Jessamyn Bowling, Tsung-Chieh (Jane) Fu, Brian Dodge, Lucia Guerra-Reyes, and Stephanie Sanders. Sexual diversity in the United States: Results from a nationally representative probability sample of adult women and men. *PLOS ONE*, 12(7):e0181198, July 2017.

[26] Cormac Herley. So long, and no thanks for the externalities: the rational rejection of security advice by users. In *Proceedings of the 2009 workshop on New security paradigms workshop*, pages 133–144, 2009.

[27] Cormac Herley. Unfalsifiability of security claims. *Proceedings of the National Academy of Sciences*, 113(23):6415–6420, 2016.

[28] Maia Jacobs, Henriette Cramer, and Louise Barkhuus. Caring about sharing: Couples' practices in single user device access. In *Proceedings of the 2016 ACM International Conference on Supporting Group Work*, pages 235–243, 2016.

[29] Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W Reeder. A" nutrition label" for privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, pages 1–12, 2009.

[30] Kami Kosenko, Geoffrey Luurs, and Andrew R Binder. Sexting and sexual behavior, 2011–2015: A critical review and meta-analysis of a growing literature. *Journal of computer-mediated communication*, 22(3):141–160, 2017.

[31] Amanda Lenhart, Michele Ybarra, and Myeshia Price-Feeney. Nonconsensual image sharing: one in 25 americans has been a victim of" revenge porn". 2016.

[32] Ada Lerner, Helen Yuxun He, Anna Kawakami, Silvia Catherine Zeamer, and Roberto Hoyle. Privacy and activism in the transgender community. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[33] Karen Levy and Bruce Schneier. Privacy threats in intimate relationships. *Journal of Cybersecurity*, 6(1), 2020.

[34] Ben Lovejoy. 'Nude' app uses coreml to automatically detect & protect intimate photos on an iphone - 9to5mac. https://9to5mac.com/2017/10/17/nude-photos-iphone/, 2017. (Accessed on 02/14/2024).

[35] Megan K Maas, Kyla M Cary, Elizabeth M Clancy, Bianca Klettke, Heather L McCauley, and Jeff R Temple. Slutpage use among us college students: the secret and social platforms of image-based sexual abuse. *Archives of sexual behavior*, 50:2203–2214, 2021.

[36] Diogo Marques, Ildar Muslukhov, Tiago Guerreiro, Luís Carriço, and Konstantin Beznosov. Snooping on mobile phones: Prevalence and trends. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, pages 159–174, 2016.

[37] Alice E Marwick and danah boyd. Networked privacy: How teenagers negotiate context in social media. *New Media & Society*, 16(7):1051–1067, 2014.

[38] Louise Matsakis. The Motherboard guide to sexting securely, 2017. https://www.vice.com/en_us/article/mb3nd4/how-to-sext-securely-safely-what-apps-to-use-sexting.

[39] Tara Matthews, Kerwell Liao, Anna Turner, Marianne Berkovich, Robert Reeder, and Sunny Consolvo. "She'll just grab any device that's closer": A study of everyday device & account sharing in households. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016.

[40] Tara Matthews, Kathleen O'Leary, Anna Turner, Manya Sleeper, Jill Palzkill Woelfer, Martin Shelton, Cori Manthorne, Elizabeth F. Churchill, and Sunny Consolvo. Stories from survivors: Privacy & security practices when coping with intimate partner abuse. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017.

[41] Jane Mavoa, Simon Coghlan, and Bjørn Nansen. "it's about safety not snooping": Parental attitudes to child tracking technologies and geolocation data. *Surveillance & Society*, 21(1):45–60, 2023.

[42] Nora McDonald and Andrea Forte. The politics of privacy theories: Moving from norms to vulnerabilities. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

[43] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23, 2019.

[44] Henry L Minton. *Departing from deviance: A history of homosexual rights and emancipatory science in America*. University of Chicago Press, 2002.

[45] Maryam Mustafa, Abdul Moeed Asad, Shehrbano Hassan, Urooj Haider, Zainab Durrani, and Katharina Krombholz. Pakistani teens and privacy-how gender disparities, religion and family values impact the privacy design space. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 195–209, 2023.

[46] Helen Nissenbaum. Privacy as contextual integrity. *Wash. L. Rev.*, 79:119, 2004.

[47] Cheul Young Park, Cori Faklaris, Siyan Zhao, Alex Sciuto, Laura Dabbish, and Jason Hong. Share and share alike? An exploration of secure behaviors in romantic relationships. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, pages 83–102, Baltimore, MD, August 2018. USENIX Association.

[48] Rizu Paudel, Prakriti Dumaru, Ankit Shrestha, Huzeyfe Kocabas, and Mahdi Nasrullah Al-Ameen. A deep dive into user's preferences and behavior around mobile phone sharing. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–22, 2023.

[49] Elissa M Redmiles, Sean Kross, and Michelle L Mazurek. Where is the digital divide? a survey of security, privacy, and socioeconomics. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 931–936, 2017.

[50] Elissa M Redmiles, Amelia R Malone, and Michelle L Mazurek. I think they're trying to tell me something: Advice sources and selection for digital security. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 272–288. IEEE, 2016.

[51] Margie Ruffin, Israel Lopez-Toldeo, Kirill Levchenko, and Gang Wang. Casing the vault: Security analysis of vault applications. In *Proceedings of the 21st Workshop on Privacy in the Electronic Society*, pages 175–180, 2022.

[52] Nithya Sambasivan, Amna Batool, Nova Ahmed, Tara Matthews, Kurt Thomas, Laura Sanely Gaytán-Lugo, David Nemer, Elie Bursztein, Elizabeth Churchill, and Sunny Consolvo. "They don't leave us alone anywhere we go": Gender and digital abuse in South Asia. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, New York, NY, USA, 2019. Association for Computing Machinery.

[53] Nithya Sambasivan, Garen Checkley, Amna Batool, Nova Ahmed, David Nemer, Laura Sanely Gaytán-Lugo, Tara Matthews, Sunny Consolvo, and Elizabeth Churchill. "Privacy is not for me, it's for those rich women": Performative privacy practices on mobile phones by women in south asia. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, pages 127–142, 2018.

[54] Emily Setty. A rights-based approach to youth sexting: Challenging risk, shame, and the denial of rights to bodily and sexual expression within youth digital sexual culture. *International Journal of Bullying Prevention*, 1:298–311, 2019.

[55] Scott Skinner-Thompson. Performative privacy. *UCDL Rev.*, 50:1673, 2016.

[56] Scott Skinner-Thompson. Privacy's double standards. *Wash. L. Rev.*, 93:2051, 2018.

[57] Julia Slupska, Selina Cho, Marissa Begonia, Ruba Abu-Salma, Nayanatara Prakash, and Mallika Balakrishnan. "They look at vulnerability and use that to abuse you": Participatory threat modelling with migrant domestic workers. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 323–340, 2022.

[58] Julia Slupska, Scarlet Dawson Dawson Duckworth, Linda Ma, and Gina Neff. Participatory threat modelling: Exploring paths to reconfigure cybersecurity. In *extended abstracts of the 2021 CHI conference on human factors in computing systems*, pages 1–6, 2021.

[59] Emily C. Stasko and Pamela A. Geller. Reframing sexting as a positive relationship behavior. Drexel University, 2015. https://www.apa.org/news/press/releases/2015/08/reframing-sexting.pdf.

[60] Samuel Hardman Taylor, Jevan Alexander Hutson, and Tyler Richard Alicea. Social consequences of Grindr use: Extending the internet-enhanced self-disclosure hypothesis. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 6645–6657, 2017.

[61] Joris Van Ouytsel, Ellen Van Gool, Michel Walrave, Koen Ponnet, and Emilie Peeters. Sexting: Adolescents' perceptions of the applications used for, motives for, and consequences of sexting. *Journal of Youth Studies*, 20(4):446–470, 2017.

[62] Noel Warford, Tara Matthews, Kaitlyn Yang, Omer Akgul, Sunny Consolvo, Patrick Gage Kelley, Nathan Malkin, Michelle L Mazurek, Manya Sleeper, and Kurt Thomas. Sok: A framework for unifying at-risk user research. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 2344–2360. IEEE, 2022.

[63] Miranda Wei, Eric Zeng, Tadayoshi Kohno, and Franziska Roesner. Anti-Privacy and Anti-Security advice on TikTok: Case studies of Technology-Enabled surveillance and control in intimate partner and Parent-Child relationships. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 447–462, 2022.

[64] Yuxi Wu, W Keith Edwards, and Sauvik Das. Sok: Social cybersecurity. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1863–1879. IEEE, 2022.

[65] Nannan Xie, Hongpeng Bai, Rui Sun, and Xiaoqiang Di. Android vault application behavior analysis and detection. In *Data Science: 6th International Conference of Pioneering Computer Scientists, Engineers and Educators, ICPCSEE 2020, Taiyuan, China, September 18-21, 2020, Proceedings, Part I 6*, pages 428–439. Springer, 2020.

[66] Xiaolu Zhang, Ibrahim Baggili, and Frank Breitinger. Breaking into the vault: Privacy, security and forensic analysis of android vault applications. *Computers & Security*, 70:516–531, 2017.

Appendix

Interview Protocol

1. Vault apps or hidden album apps are tools that store content on a mobile device so people won't accidentally see it. What vault apps or hidden album tools do you use or have used?

2. For each tool: When did you start using the tool?

3. Why did you start using it? a. Was there a specific incident or story that prompted getting it? b. Did anyone else's opinion affect your decision on it?

4. How did you learn about this tool? a. How did you pick that tool over others? b. If they use multiple tools, ask about both of them

5. Did you have any concerns or questions about using it?

6. If we don't know if it's free: Did you have to pay for the tool? What led to this decision?

7. What phone do use?

If participant stores intimate media and stores other types of content, ask the "non-intimate media" line of questioning first, and based on time left get as far with the "intimate media" line of questioning as possible. If decision to use involves storing intimate media:

1. What types of images do you store? a. If storing person's images: Whose images do you store?

2. How did you store them before getting this tool?

3. How do you store them with this tool? a. E.g., directly saving, screenshotting, etc.

4. Do you plan on storing this media indefinitely? a. Is there any scenario you could imagine where you would delete the media?

5. If storing person's images: For the person in the photos you've saved, what expectations did you and that person have when they sent you the photo?

6. Do you ever share these photos? a. Does this app have sharing functionality?

7. What concerns do you have that prompted you to use this tool?

8. You mentioned x concerns prompting you to use this tool. Do you feel like this tool addresses all of these concerns? a. Why or why not?

9. Are there other features of this tool that you find useful? a. Why or why not? b. Can you provide an example of a useful feature? c. Are there features you want that they don't provide?

10. Do you have any concerns that this tool doesn't address?

11. Is there anything you'd like to see changed in the app to help prevent these concerns?

12. Do you have any concerns with the tool itself?

13. What settings do you feel comfortable opening the tool?

14. Is there a way to access these files through your computer?

If decision to use involves storing non-intimate media:

1. What kind of files do you store?

2. What concerns do you have that prompted you to use this tool?

3. You mentioned x concerns prompting you to use this tool. Do you feel like this tool addresses all of these concerns? a. Why or why not?

4. Are there other features of this tool that you find useful? a. Why? b. Can you provide an example of a useful feature? c. Are there features you want that they don't provide?

5. Do you have any concerns that this tool doesn't address?

6. Is there anything you'd like to see changed in the app to help prevent these concerns?

7. Do you have any concerns with the tool itself or the company?

8. In what settings do you feel comfortable opening the tool?

9. Do you plan on storing this media indefinitely? a. Is there any scenario you could imagine where you would delete the media?

10. How did you store them before getting this tool?

11. Is there a way to access these files through your computer?

Final Questions (for every participant)

1. Ask these questions per each type of media stored: What kind of people are you concerned about seeing x content? a. Assuming its for privacy or security usage: Are there outcomes you are concerned about if x sees that content? i. If yes: What are the outcomes?

2. Have you tried any other way of preventing that from happening besides vault apps?

3. Do you feel like the vault app has successfully prevented that?

4. Is there anything you'd like to see changed in the app to help prevent that?

Demographics
You can say pass if you want to skip any of these questions.

1. Relationship style? E.g., monogamous, polyamorous, etc.

2. Relationship status

3. Sexual orientation

Anything else you want to tell us that we haven't asked?

App Affordance Tables

List of Apps	PIN?	Biometrics?	Swipe Pattern?	Selective Access (different "accounts")	Steps to Access
Snapchat My Eyes Only	4 digits	N	N	N	2 swipes
iOS Photos Hidden Album	Screen Lock		N	N	2 clicks
Google Photos (hidden album)	Screen Lock			N	3 clicks
Dropbox (locked file)	Password (Paid version only)	N	N	Y - select who has access	2 clicks
Secret Photo Album	Y	N	Y	N	Open + unlock
Secure Folder (Android Files)	4 digits	N	Y	N	2 clicks + unlock
Photo Vault (KeepSafe)	4 digits	Fingerprint	Y	decoy vault (2nd PIN) No seperate accounts	Open + unlock
Vault	15 digits	Fingerprint	Y	decoy vaults (PAID ONLY)	Open + unlock
Adobe Acrobat (locked file)	Password (Paid version only)	N	N	N	Open + unlock each file
App Lock	N	N	Y	Y - "Profiles"	Open + unlock
iOS Notes (locked file)	Y	Y screen lock	N	N	Open + unlock
OneNote	Password	N	N	N	Open + unlock

Table 5: Authentication and access-related affordances in vault apps and hidden folders participants mentioned. Screen lock authentication refers to the app requiring the same type of authentication as for unlocking the phone. Swipe pattern refers to graphical passwords. Selective access refers to either account-based access permissions or to PIN-based access to different file folders, i.e., providing a different PIN leads to different storage on a vault app. Steps to Access refers to opening a vault app or hidden album from the step prior to authentication; e.g., if one is opening a Hidden Album from iOS Photos, we assume they already have the app Photos open.

List of Apps	Independent from OS	Discreet Icon / Within App	Hide Content In-App	Hidden from Memories	Folders	Share / Download Content
Snapchat My Eyes Only	Y	Snapchat	N	Y	N	Share
iOS Photos Hidden Album	N	Photos	N	Y	N	Y
Google Photos (hidden album)	Y	Photos	N	Y	N	N
Dropbox (locked file)	Y	Dropbox	N	N/A	Y	Share / Download
Secret Photo Album	Y	Y	N	N/A	Y	Y
Secure Folder (Android Files)	Y	Files	N	N/A	N	N
Photo Vault (KeepSafe)	Y	Y	Y - decoy vault with second PIN	N/A	Y	Share
Vault	Y	"Stealth Mode"*	Y - decoy vault with second PIN	N/A	Y	Y
Adobe Acrobat (locked file)	Y	Acrobat	Y - per file passwords	N/A	Y	Y - keeps password
App Lock	Y	N	N	N/A	N	Share
iOS Notes (locked file)	N	Notes	Y	N/A	N	Y
OneNote	Y	OneNote	Y - per section passwords	N/A	Y	Y

Table 6: Discreteness and organizational-related affordances in vault apps and hidden folders participants mentioned. Distinct from Device refers to an app not being a default installation with the operating system. Discreet Icon refers to having a camoflaging icon, such as that of a Calculator app, and Within App refers to the tool being hidden within a non-vault app. Hide Content In-App refers to . Hidden from Memories refers to removing authentication-protected photos from app-curated photo collections. Folders refers to being able to make separate folders for files. *Doesn't work on modern phones.

"I do (not) need that Feature!" – Understanding Users' Awareness and Control of Privacy Permissions on Android Smartphones

Sarah Prange[1], Pascal Knierim[2], Gabriel Knoll[3], Felix Dietz[1], Alexander De Luca[4], Florian Alt[1]

[1]*University of the Bundeswehr Munich, Germany, {firstname.lastname}@unibw.de*
[2]*University of Innsbruck, Austria {firstname.lastname}@uibk.ac.at*
[3]*LMU Munich, Germany, {firstname.lastname}@campus.lmu.de*
[4]*Google Munich, Germany*

Abstract

We present the results of the first field study ($N = 132$) investigating users' (1) *awareness* of Android privacy permissions granted to installed apps and (2) *control behavior* over these permissions. Our research is motivated by many smartphone features and apps requiring access to personal data. While Android provides privacy permission management mechanisms to control access to this data, its usage is not yet well understood. To this end, we built and deployed an Android application on participants' smartphones, acquiring data on actual privacy permission states of installed apps, monitoring permission changes, and assessing reasons for changes using experience sampling. The results of our study show that users often conduct multiple revocations in short time frames, and revocations primarily affect rarely used apps or permissions non-essential for apps' core functionality. Our findings can inform future (proactive) privacy control mechanisms and help target opportune moments for supporting privacy control.

1 Introduction

For many years, the decision of what data is collected, processed, and potentially shared with third parties had been the sole decision of the app or service provider, with many Android apps requesting more permissions than necessary in the past [38]. Users unwilling to share the requested data could only make a simple choice – installing or not installing the app or service. More recently, a trend can be observed towards designing apps and services in a more privacy-preserving way. An example is providing users more control by allowing one or multiple permission(s) to be modified (granted/revoked) during use. Additionally, *runtime permissions*, introduced in Android 6.0, allow apps to request permissions when needed.

Empowering users to manage privacy permissions creates several challenges, most importantly scalability. The number of apps/services and diverse data sources make it hard for users to stay aware of which data is collected by whom and make permission settings suit their needs and purposes.

Researchers tried to tackle this challenge by a) making privacy information more easily accessible to inform decisions (e.g., [54, 56]) and b) providing users support to take control over privacy choices. For instance, the concept of *privacy assistants* helps users make privacy choices based on their preferences [30, 48]. Another example is the *Privacy Dashboard* introduced in Android 12, which provides users a quick overview of which permissions are granted to which service or application and the auto-revoke feature that removes access to unused permissions. At the same time, there is currently little knowledge of the degree to which people are aware of such privacy permission management mechanisms; if so, how they use them; and how effective these mechanisms are in terms of supporting users in making informed privacy choices, in particular as they change permissions of apps after installation. However, such knowledge is valuable to enhance existing or design novel privacy permission management approaches that better support this post-installation or post-first use update of permissions. We address this through the first in-situ field study, gathering users' privacy permission behavior in an uncontrolled environment over a two-week period.

The following two questions drive our research:

RQ1 – Awareness. Are users aware of a) privacy permissions granted to installed apps and b) current interfaces to manage (and revoke) permissions?

RQ2 – Control. How often, when, and why do users grant, deny, and revoke privacy permissions?

To answer these questions, we conducted a study with Android smartphone users ($N =132$), primarily young Europeans, consisting of two parts: first, our study app acquired current apps and permission settings of participants' phones, allowing us to analyze which privacy permissions they had *initially granted* or *initially denied* for their installed apps; second, our app monitored participants' devices for two weeks for

permission updates to investigate if, for example, participants *revoked* or *later granted* a particular permission. We complemented this data using the Experience Sampling Method [83]: permission updates triggered notifications redirecting users to in-app questionnaires, asking for reasons for their decision.

We found that revocations affect privacy permissions that users consider sensitive (e.g., access to stored files or the camera), but only if this does not affect an app's core functionality or intended use case. Moreover, several updates were often conducted in short time frames, indicating opportune moments exist when users are willing to work on their privacy choices. Our findings provide a better understanding of users' current privacy control and can help to design future mechanisms to support users (proactively) in doing so.

Contribution Statement. Our contribution is twofold. Firstly, we contribute an in-depth investigation of privacy permission awareness and control (i.e., grant/revoke actions) in the real world among 132 Android users. In particular, we collected (1) data on actual privacy permission states at the beginning of the study as well as (2) data on permission changes, along with experience sampling data over the course of two weeks. Secondly, we discuss how our findings can inform future user interfaces supporting privacy permission management.

2 Related Work

We draw from several strands of related work. We illustrate users' privacy awareness and perceptions towards data available on mobile devices and highlight the usefulness of mobile privacy control interfaces. We focus on Android permissions as iOS apps are generally encrypted, and no publicly available analysis tools exist [58]. Hence, an in-situ exploration of users' permission behavior is impossible on iOS.

2.1 Users' Privacy Awareness & Perceptions

Privacy preferences and concerns about sharing data are highly individual [29, 36, 81] and based on contextual factors. For instance, the type and purpose of a specific device as well as the frequency of data being collected [28,29,60,61,89], along with who collects the data [39, 59] and how long it is stored [36], impact users' willingness to share personal data.

Users are particularly concerned about cameras and microphones, as these can capture sensitive data [18, 27, 29, 57, 60, 69]. However, current mobile phones provide an increasing number of sensors that can likewise capture sensitive data. Examples include but are not limited to, GPS sensors allowing users' location to be inferred or gyroscopes allowing users' physical activity to be derived. Other examples of data available on mobile devices include users' personal files, location, and communication data, all of which are considered sensitive data [49]. Users are also specifically concerned about access to their text messages, e-mails, photos, and contacts [39].

At the same time, users are often unaware of which sensors are active on their mobile devices [49] and which data

is collected by apps running on their devices [21]. Moreover, textual descriptions of permissions can be misleading in terms of actual permissions being required [33], and permissions are often requested for third-party libraries rather than apps' core functionality [37]. Specific privacy implications of certain personal data being exposed thus remain unclear to users [23]. Consequently, it is challenging for users to adequately assess which service or functionality currently has access to a specific sensor, let alone the concrete privacy implications of sharing this data. Modern smartphones offer visual cues through hardware and software, such as the microphone and camera indicators, to address this. However, users struggle to understand how much personal information can be gained from smartphone data. In particular, while access to e-mails discloses sensitive information, users underestimate this as a threat [34]. Lastly, users also tend to sacrifice privacy preferences for personal needs [19,44,68] (e.g., if access to a certain sensor would enable a certain feature) or are unaware of the extent to which their personal data is being collected [19].

Increasing privacy awareness, for example, through simple means like microphone indicators, is a prerequisite for users to be able to take control over their personal data and ultimately act according to their privacy needs [29, 65, 66].

2.2 Mobile Privacy Control Mechanisms

Users mostly wish to stay in control over their data [20, 27, 66, 81]. Current privacy interfaces aim to support this.

2.2.1 Designing for Mobile Privacy Control

The default approach to gathering users' consent before data collection is notice and choice [32, 41, 75, 78]. However, privacy notices are often of poor usability [74], and, thus, insufficient [32]. To address this, researchers proposed privacy notices to be visually appealing [56] and privacy choices to be designed meaningfully and accessible [41].

Current privacy control is oftentimes non-accessible [29, 47], either overly reduced [41] or too complex [20, 43, 47], or overwhelming [79]. Moreover, the number of permission requests is rising: more permissions are requested than necessary [38], and requests are made for third-party libraries rather than core functionality of apps [33]. Tahaei et al. shed light on the developers' perspective: developers are oftentimes unsure about the scope of permissions and, thus, tend to request multiple permissions for smooth functionality of apps [82].

Researchers tried to support users in re-gaining control over their personal data while at the same time reducing the number of decisions to be made [76]. Personalized privacy assistants, for instance, assess users' privacy preferences automatically to make personalized recommendations on privacy settings [30,48]. Considering contextual factors, e.g., the purpose of a specific permission request can improve such recommendations [79]. SmarPer learns from users' decision patterns to automate runtime permissions [71]. Also, repetitive privacy decisions could be automized [80] to reduce users'

decision burden. For mobile applications, the "Privacy Facts" display can help users better understand to-be requested privacy permissions and thus make more informed decisions for apps requiring less privacy intrusive permissions [54].

Prior research also showed that more restrictive privacy policies can increase users' willingness to share data [62]. Possible decision and control support could thus include which data is collected, where and for how long it is stored, and with whom it is shared [62]. Moreover, information on data accessed without actively using the app, data transmission, and app ratings can help users make informed decisions about privacy permissions [77]. Also, as users tend to base their privacy concerns on previous (potentially bad) experiences, privacy choices might be designed to be personal and concrete [55]. Other approaches include the automated analysis of requested permissions [33], respective textual descriptions [37], or users' comments [50] to help assess the actual need for requested permissions and identify undesired app behaviors. Lastly, privacy permission could be requested *proactivly* when access is actually necessary, for example, contextually choosing permissions relevant enough to prompt users directly, similar to Android's runtime permissions. Other permissions could be defined once during setup [64].

Android Privacy Permissions Android implements privacy control via *permissions* [8]: app developers have to gather users' consent before accessing specific sensors or data (for example, location or stored files). This is typically done by a request prompt: users can choose to *accept* or *deny* access. For Android apps, privacy notices and permission requests typically appeared *upon app installation*. While being recognized by users, these install-time permissions were rarely understood, thus limiting users in making informed privacy decisions regarding whether to install a certain app [40, 53]. With the shift to *runtime permissions* [13] from Android 6.0, permissions are only requested when needed first, providing users with additional contextual information. This allows users to decide whether specific permission is necessary and to revoke decisions later [25]. This contextual approach also benefits developers as grant rates increase [35]. In addition, from Android 11 on, users have more control over the location, microphone, and camera permissions. Moreover, permissions can be granted for *one time* only, and permissions are *auto-revoked* for unused apps [9]. Other permission models have also changed significantly. For instance, access to users' photo library is now limited through the *Photo Picker* [11], meaning apps only have access to specific photos the user selects.

Android Privacy Interfaces To summarize current permission states, Android's *Permission Manager* lists permission types along with apps that currently do or do not have access to these. With Android 12.0, the *Privacy Dashboard* (see Appendix B) was introduced to provide users with a detailed overview of which applications currently have access to which sensors, along with means to grant or revoke this access [1].

2.2.2 Understanding Users' Mobile Privacy Choices

To design privacy interfaces, understanding users' current use of privacy control is crucial to support them in future choices. In an online survey, Frik et al. found that many users are unaware of privacy permission settings available on their smartphones and have not actively changed them due to a perceived lack of expertise or low self-efficacy [42]. Once granted, users rarely revoke third-party access to personal data (e.g., fitness data) – either because they are unaware of the permission previously granted or they are unaware of the option to revoke access post-hoc [90]. At the same time, strict privacy settings might negatively affect apps' usability [51]. Looking into Google's single sign-on system, Balash et al. showed that users are concerned about giving third-party apps access to personal information but less concerned about access to calendars, emails, or cloud storage [24].

While Android's *runtime permissions* allow users to assess whether or not an application needs specific access by putting them in context, most such permission requests are still accepted, with exceptions mainly for microphone and calendar access. When denying permissions, users mainly believe an app should not need certain permissions or would work without them. In contrast, for granting permissions, access to features and trust are dominant reasons [25]. Bakopoulou et al. found that users oftentimes cannot adequately assess the implications of their private information being exposed to mobile applications [23]. More recently, Cao et al. identified factors impacting privacy decision-making among 1,719 users of Android versions 6.0 to 10. Users were likelier to deny permissions requests they did not expect and less likely to deny permissions that came with explanations [26]. Tahaei et al. found that end-users grant permissions as they desire a certain functionality or trust a certain app [82]. To minimize the number of user decisions, Liu et al. [63] suggest a privacy assistant that automatically configures app permissions based on an initial privacy assessment.

2.3 Summary

The number of apps on users' smartphones makes it challenging for them to be aware of and control their personal data being collected and shared. This challenge is exacerbated as many apps request more permissions than necessary for the core functionality it provides [33, 38, 82]. At the same time, users' awareness and comprehension of, as well as the possibility to revoke a decision previously made, are essential components for the usability of privacy choice mechanisms [46]. Newer Android versions tackle this challenge by providing users with a) *runtime permissions* (since Android 6.0), which gives users more context to form a privacy decision [25]; b) an overview of current permission states per app and control options (*Permission Manager*, followed by the *Privacy Dashboard* on Android 12); and c) privacy indicators visualizing current access to sensors (since Android 12).

Prior work investigated users' general privacy perceptions towards mobile apps [19, 21, 49], privacy permission behavior resulting from the runtime permission dialogs [25], and recently, users' privacy control behavior using surveys [23, 42, 90] or one-time collection of permission states [22]. We add to this knowledge by contributing an in-situ investigation of users' a) *awareness* of built-in privacy control interfaces and permission states and b) *permission control* (e.g., revoking permission that was initially granted or later granting permission that was initially denied) on current Android versions by collecting in-the-wild data over a period of *two weeks*. We gather those *in-situ* insights by implementing an Android app that collected information on installed apps and permission states, as well as on updates to these. We complement our data using Experience Sampling (ESM) [83].

Our approach is in line with prior research on privacy permissions. Field studies have generally been used to understand the contextual nature of permission granting decisions [85, 88], and for automating permission management [86, 87]. ESM as data collection method was effectively applied in prior privacy studies among Android users [22, 25] to capture their privacy behaviors [26], yet did not focus on post-hoc privacy management, including revoking permissions.

3 Research Approach

Using the Experience Sampling Method (ESM) [83] and automated data logging using an Android application, we collected data on users' *awareness* of current privacy permissions states (RQ1) and updates of privacy permissions (*control*, RQ2) among 132 participants. Following van Berkel's suggestion for ESM-based studies using smartphones [83], we decided on a two-week period. This also provided enough time to observe a substantial number of permission updates. Note that with this approach, we aimed to identify general permission management behavior rather than generalizing our findings to the broader population.

3.1 Apparatus

We built an Android app for version 8.0 to 12.0 (the latest version at the time of the study) in Kotlin 1.6.20 [15], thus covering 86.7% of Android users. The app comprises two major components: the *Permission Scanner* and the *In-App Experience Sampling (ESM) Questionnaire Interface* (see Figure 1, right). The Permission Scanner regularly monitored participants' devices for permission states of all installed apps (every two hours, excluding system services and apps with zero usage time). For this, our application requested access to Android's Package Manager [16] and Usage Stats Manager [17]. In case at least one permission *update* (i.e., change in permission compared to the last scan) was detected, an ESM questionnaire was triggered, asking for reasons for up to five permissions updates, depending on the number of updates. The In-App Experience Sampling Questionnaires were implemented using *SurveyKit* [14].
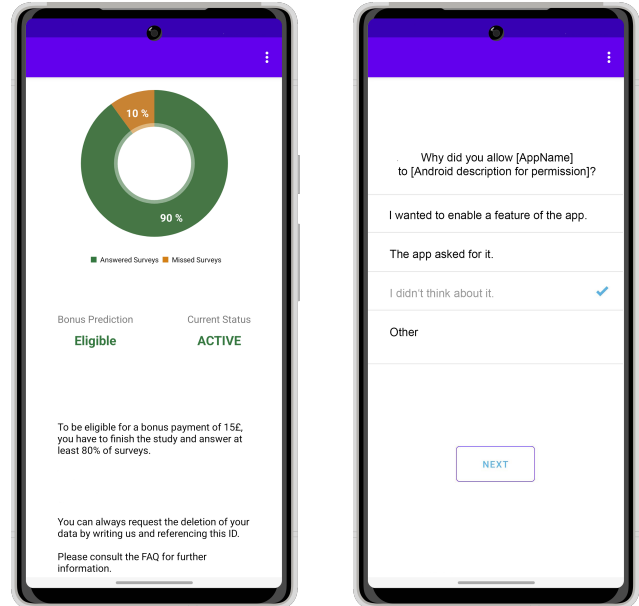
Figure 1: Study App Screenshots. Left: The home screen provides participants with an overview of answered/missed experience sampling questionnaires, eligibility for the bonus payment, and access to contact information. Right: A sample screen with an experience sampling questionnaire.

The app's home screen provided participants with an overview of answer statistics using MPAndroidChart [7] and access to contact information and frequently asked questions (see Figure 1, left). Data was stored in a Firebase Realtime Database [5]. We also used Firebase Crashlytics [4] to analyze and account for any errors during the study. The app was made available to participants using Firebase App Distribution [3].

3.2 Collected Data

Our application collected data through automated logging and questionnaires. Participants were asked to answer a questionnaire at the beginning (*initial questionnaire*), after one week (*mid-term questionnaire*), and at the end (*final questionnaire*) of the study. In addition, participants were asked to answer two types of experience sampling questionnaires: a questionnaire on control (*ESM control questionnare*) and one on awareness (*ESM awareness questionnaire*).

3.2.1 Automated Data Logging

Our application automatically collected the following data: information on the device (device name, brand, and Android version); an initial list of all installed applications along with usage duration; state of privacy permissions upon installation; and privacy permission updates during the study.

Permission States & Updates For privacy permissions of apps, we logged their state at the beginning of the study (i.e., *initially granted* or *initially denied*), and every two hours over two weeks. If a permission state changed during the study (i.e., a different state than the previous scan), we recognized

this as a *permission update*. We consider updates from previously granted permissions to denied access (i.e., *revoked*) and updates from initially denied permissions to granted permissions (i.e., *granted later*). Update data includes app names, requested permissions with the current state, and app usage time. Our data might include permission changes resulting from a) newly installed apps or b) Android's auto-revoke or one-time permission features (from Android 11 on [9]). Note that for a), users' (active) privacy decisions, as made when first using a new app, are included in our data. For b), we acknowledge that some updates might have been initiated by Android rather than consciously by users (see Section 4.4.4).

3.2.2 Experience Sampling (ESM)

We utilized ESM, prompting participants with in-situ questionnaires via notifications [83]. Our app administered two types: (1) upon detected permission *updates* asking for reasons (ESM control questionnaire) and (2) asking about permission states of certain apps *daily* (ESM awareness questionnaire). We covered all permissions updated within the respective time frame for the ESM control questionnaires. Answer options included sample reasons (see Appendix C.3.1) and an option for free text. These options resulted from discussions among the authors to reflect the research questions. We always presented them in the same order to ensure consistency.

For the ESM awareness questionnaire (see Appendix C.3.2), we randomly chose up to five installed apps that operated in the foreground at least once since installation and required access to at least one permission. We did not give participants the correct answers (i.e., permission states). We included attention checks such as "If you read this, please select 'No'". To increase motivation, participants received clear information on the study goal and additional compensation for active engagement with the ESM questions [84]. Moreover, participants were asked to use their personal devices and could set a custom time span per day in which ESM questionnaires were sent [83]. All ESM questionnaires were withdrawn after a certain timespan (control: after 2 hours, awareness: after 12 hours) to ensure in-situ answers [83].

3.2.3 Questionnaires

We complemented our data collection with an initial and final questionnaire on users' perception of privacy permissions and a midterm questionnaire on using Android's privacy management tools (see Appendix C). Participants were to choose permissions for which they wanted to be particularly alert (*awareness*). The midterm questionnaire covered prior usage of Android's Privacy Dashboard and Permission Manager, depending on participants' Android version (*control*). This questionnaire was designed to hint users to these interfaces and see if their behavior would change in the second week of the study. We validated the clarity of all questionnaires in a pilot run, where all co-authors and research group members tested the app for two weeks, giving continuous feedback.

3.3 Procedure

Participants used our Android application over two weeks. The detailed procedure was as follows (see Figure 2):

1) **Installation & Setup.** Participants who agreed to participate first downloaded our application. Participants were prompted to consent to the study's procedure and privacy policy upon installation. After consent, the app collected information on the device, installed apps, and current permissions of apps along with usage duration.
2) **Initial Questionnaire.** Participants then answered an initial questionnaire covering their privacy preferences before the study (see Appendix C.1 for a full list of questions). After this questionnaire, our app started the automated data logging (permission updates) and experience sampling.
3) **Experience Sampling Phase.** For two weeks, the app scanned participants' devices for permission updates. Upon change, the app would trigger a questionnaire (via a notification), asking for the reasons for later granting or revoking that specific permission (ESM control questionnaire, see Appendix C.3.1). In addition, the app asked daily about permission states of a random selection of apps (ESM awareness questionnaire, see Appendix C.3.2).
4) **Mid-Term Questionnaire.** After a week, participants filled in a mid-term questionnaire on using Android's current privacy interfaces, asking them to visit the *Permission Manager* and/or *Dashboard* afterwards (Appendix C.2).
5) **Final Questionnaire.** The final questionnaire repeated the initial questions on privacy perceptions (Appendix C.1).

3.4 Recruitment & Requirements

We recruited our sample via Prolific, an online subject pool [12, 72]. We enforced several requirements through pre-screening: (1) Participants must be fluent in English. (2) The sample should be equally balanced in terms of gender and only include users aged 18 or above based on their demographic characteristics (see [6] for details on balanced samples). (3) We sampled participants residing in Europe, Canada, the USA, and Australia. We did so to reduce effects from, e.g., smartphones being shared among family members, people tending to use multi-purpose apps (WeChat in China), or cases in which vendors pre-install apps (many countries in Africa.).

Participants were required to use the app on their personal smartphones with Android versions 8 to 12.0. Through Prolific, participants installed and set up the app. Upon setup completion ($N = 300$, 14 minutes on average, according to Prolific), participants were reimbursed with 1.9 GBP on average[1]. For participants following the study over two weeks and answering at least 80% of the ESM questionnaires, we paid a bonus of 15 GBP (average time commitment 56 minutes, based on the total usage time of the study app). The study was conducted between April and May 2022.

---

[1] The average hourly wage was 7.62 GBP as suggested by Prolific.
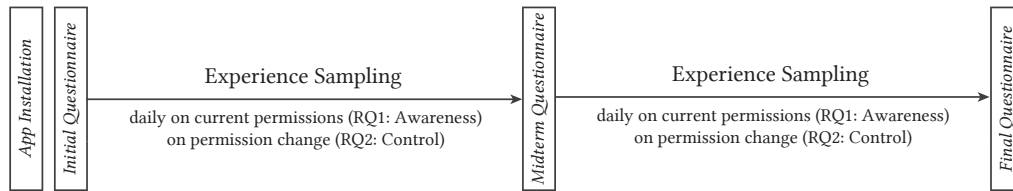
Figure 2: Study Procedure: Participants first installed the application and filled out an initial questionnaire. In the following experience sampling phase, participants were asked about active permission revocations (control) and awareness of current permissions. They filled out a midterm questionnaire after one week and a final questionnaire after two weeks.

3.5 Ethical Considerations

In the country where this research was conducted, formal IRB approval is not required for this type of human subject research [70]. However, we comply with all university ethics regulations and national data protection regulations. Consent was gathered as follows. First, participants read the study description and then accessed the study through Prolific [12]. Second, they were directed to Firebase App Distribution [3], where they consented to downloading and installing the app[2]. Third, we gathered participants' informed consent through our app before collecting data. We stopped data collection automatically after two weeks and suggested uninstalling our app. Collected data comprises an app list, permission settings changes, and questionnaire answers. All data was collected anonymously using randomly generated identifiers. As such, we are unable to identify individual participants or devices. Through Prolific, we only recruited participants with a minimum age of 18. We followed Prolific's suggestion for reimbursing the *study's initial setup procedure*, which took 14 minutes on average. For participating over 14 days with a daily effort of around 4 minutes, we paid a bonus of 15 GBP.

3.6 Limitations

Android Versions Our study is limited to participants running Android version 8 and above. This excludes older versions but ensures compatibility and access to the *Permission Manager*. In a few cases, the app showed unforeseen behavior, leading to the exclusion of some participants (see Section 3.7).

Sample Our participant sample is biased towards young users (mean age 26.45) from European countries with Android versions below 12.0. Thus, our results might not apply to the general population or future Android versions.

Selection Bias The initial and midterm questionnaires and participation in our study, in general, might have influenced permission control behavior. Still, we a) wanted to be transparent about the study goal, not using any deception, and b) see if knowledge about Android's privacy tools (midterm questionnaire) influences users' behavior. The study advertisement and task did not explicitly require participants to engage with permission management actively but only to answer ESM questionnaires. The initial questionnaire deliberately did not

hint at privacy management but focused on permissions' general importance. We did not find significant differences in the number of permission updates in the study's first vs. second week (before/after the midterm questionnaire).

We acknowledge that, due to self-selection, participants may have had fewer privacy concerns than the average population. Generally, self-stated privacy preferences (as in our questionnaires) tend to differ from actual behavior (cf. the "privacy paradox" [44]). Our results include logging data on actual privacy permission states to account for this.

User vs. System-initiated Updates Sixty-six participants (on Android 11 or 12) could grant permissions for camera, microphone, and location for *one time* [10] only, and permissions might have been revoked automatically for unused apps (cf. auto-revoke [9]). We could not actively capture these cases (see Section 3.2.1), but found the number of such possible cases through post-hoc analysis (see Section 4.4.4).

App (Un)Installs Our analysis considers permissions granted to newly installed apps during the study as these result from conscious user decisions. We did not consider uninstalls as permission changes because we do not know the reasons. We acknowledge privacy concerns, similar to those leading to permission revocation, might have been the reason.

3.7 Data Cleaning

The setup and app installation were completed by 300 participants. Of these, 179 completed the full study, with 158 participants answering at least 80% of the ESM questionnaires and, thus, receiving the bonus. Of these, we excluded 13 participants based on corrupt or missing data, and 13 participants based on app crashes, failed attention checks, or unknown Android versions. Ultimately 132 samples were analyted.

During the study, we collected answers for a total of 366 ESM control questionnaires (2.77 on average per participant), 885 ESM awareness questionnaires (6.7 on average per participant)[3], and initial, midterm, and final questionnaires. Note that from a few participants, we received more than one answer set for the same questionnaire. In these cases, we considered the first complete set of answers for analysis.

---

[2]Note that participants opted out during the first or second step.

[3]Notifications for all ESM questionnaires were withdrawn after a certain time. Hence, questionnaires may have remained unanswered. We did not enforce receiving one ESM awareness questionnaire per day from every participant. This practice follows Berkel et al. to ensure in-situ answers [83].

3.8 Participants

Participants were 18 to 54 years old (*Mean* = 26.45, *SD* = 6.95). 65 participants identified as women, 63 as men, and four as non-binary. Participants' nationality was mostly Polish (*N* = 40), Portuguese (*N* = 26), Italian (*N* = 22), or Greek (*N* = 14). Others were Spanish (*N* = 6), Czech (*N* = 4), British (*N* = 3), and of other mostly European nationalities (see Appendix D.1). All participants were fluent in English.

Most participants were employed full-time (*N* = 38), unemployed (and job seeking, *N* = 32), employed part-time (*N* = 26), not in paid work (*N* = 5), starting a new job within the next month (*N* = 3), or other (*N* = 27). One participant's employment data expired. Most participants completed a high school diploma (*N* = 52) or undergraduate degree (*N* = 43). 23 received a graduate degree, and few other educational levels were mentioned. Regarding their technical background, the fact that they were active on prolific and 126 participants (70.4%) were aware of the possibility of revoking permissions hints at solid technical knowledge.

4 Results

Overall, participants had 15 to 202 installed applications (*Mean* = 99.42, *SD* = 34.43) with 36,904 granted permissions and 40,175 denied permissions in total (see Table 1). Throughout our study, we acquired *permission updates* among 128 participants (2,866 updates in total, thereof grants: 1,064, revokes: 1,802, see Tables 5 and 4 for reasons). In addition, participants answered a total of 885 ESM awareness questionnaires (RQ1, 6.70 on average per participant) and 366 ESM control questionnaires (RQ2, 2.77 on average per participant).[4] Participants were somewhat aware of current permission states with 49% correct answers for granted permissions, and 34% correct answers for denied permissions (see Table 3). In the following, we present detailed results of our automated data logging and experience sampling.

4.1 App Usage

Upon installation, our study application acquired initial information on participants' Android devices and installed applications, along with initial permission states.

Android Devices & Versions   Most participants used Android 11 (*N* = 46) or 10 (*N* = 43), some used Android 12.0 (*N* = 20) or 9 (*N* = 18), and a few participants used Android 8.0.0 or 8.1.0 (*N* = 5). Hence, all participants had access to the *Permission Manager* and 20 to *Privacy Dashboard*. Device brands mainly included Xiaomi/Redmi (*N* = 53), Huawei (*N* = 25), and Samsung (*N* = 20).

Installed Apps & Permissions   Participants initially had 15 to 202 apps installed (*Mean* = 99.42, *SD* = 34.43, see Table 6 for most used apps). One app requested 0 to 22 permissions (*Mean* = 5.87, *SD* = 4.41, total number of all

---

[4]Note that questionnaires may have remained unanswered (Section 3.7).

Table 1: Overview of Initial Permission States: List of all permissions available for users to choose on Android, along with their state (i.e., denied vs granted) at the beginning of our study. Values shown represent the ratio of the permission being denied/granted in relation to the total number this specific permission was requested among all applications all participants had initially installed (total number of permission requests: 77,079, granted: 36,904, denied: 40,175).

Permission Name	#req.	Denied (%)	Granted (%)
read phone numbers	731	83.99	16.01
get accounts	4139	64.89	35.11
access background location	1771	63.13	36.87
Bluetooth scan	1490	62.55	37.45
camera	6342	58.33	41.67
read calendar	1534	57.24	42.76
record audio	4557	55.63	44.37
read phone state	4909	54.94	45.06
access fine location	5905	51.96	48.04
write calendar	1228	51.95	48.05
access coarse location	6445	51.34	48.66
read contacts	5120	51.07	48.93
write external storage	8575	50.24	49.76
read external storage	9792	48.72	51.28
access media location	1384	47.47	52.53
write contacts	2077	33.37	66.63
activity recognition	657	32.27	67.73
query all packages	2891	31.75	68.25
read SMS	1237	21.18	78.82
read call log	1327	19.67	80.33
body sensors	176	18.75	81.25

permissions: 77,079, thereof initially granted: 36,904, initially denied: 40,175). Of these, *read phone numbers* was mostly denied (83.99%), followed by access to *accounts* (64.89%), *background location* (63.12%), and Bluetooth scan (62.55%). Participants were also somewhat strict about *camera* (58.36% denied), *calendar* (57.24% denied), and *audio* access (55.63% denied). In contrast, access to *body sensors* (81.25%) was granted most, followed by reading the *call log* (80.33%), and *SMS* (78.82%). A reason for access to body sensors being often granted is that apps likely required those permissions to enable their main functionality; for example, smartwatches running WearOS require access to body sensors. Overall, participants had 47 to 540 *initially granted* permissions (*Mean* = 304.37, *SD* = 91.71) and 37 to 543 *initially denied* permissions (*Mean* = 279.58, *SD* = 95.78). Table 1 summarizes permission states when the study started.

4.2 Android Usage

The following results describe users' awareness of Android privacy permissions and their knowledge and use of Android's privacy interfaces (control).

4.2.1 Awareness of Privacy Permissions

We captured users' wishes for awareness of specific permission types at the beginning and end of our study. Looking at the data collected at the beginning, participants wished to be particularly informed about the following permissions: camera (*N* = 121), location (*N* = 118), and microphone (*N* = 116,

---

Table 2: Awareness of Privacy Permissions: Number of participants who particularly wished to be informed of the following privacy permissions (assessment in initial vs final questionnaire, respectively). Change is normalized by participants ($N = 132$). Participants could choose multiple permissions.

Permission Name	Initial	Final	Change
📷 camera	121	124	+2.3%
➹ location	118	125	+5.3%
🎤 microphone	116	120	+3.0%
☰ phone numbers	114	114	0.0%
👥 contacts	111	117	+4.5%
💬 SMS	104	107	+2.7%
🗄 external file storage	100	100	0.0%
📞 call history	99	112	+9.8%
▢ installed apps	88	86	-1.5%
👤 other users on the smartphone	77	75	-1.5%
❋ Bluetooth	74	71	-2.3%
🏃 physical activity	71	69	-1.5%
🚶 body sensors	69	72	+2.3%
📅 calendar	56	54	-1.5%
none of these	21	5	-12.1%

see Table 2 for details)[5]. After the study, the picture is similar, with a slight increase in numbers, which can probably be attributed to raised awareness due to the increased exposure to permissions during the study: location ($N = 125$), camera ($N = 124$), and microphone ($N = 120$, see Table 2 for details).

4.2.2 Knowledge of Android Privacy Interfaces

The *midterm questionnaire* revealed that the majority of participants ($N = 125$) were aware of the fact that they can post-hoc revoke permissions ($N = 7$ stated "No"). Many participants stated to have revoked a permission before ("Yes": 92, "No": 33), mainly due to privacy concerns ($N = 59$), a feature not being used anymore ($N = 51$), or security concerns ($N = 32$). Only one participant did not think about it, and one stated "other". Most participants ($N = 108$) mentioned to have engaged with the *Permission Manager* before. However, they typically used it less than once a month ($N = 84$, "At least once a month": 21, "At least once a week": 3). 15 participants had not used the *Permission Manager* before, and 9 did not know about it. Of Android 12 users ($N = 20$), only a few ($N = 7$) had used the *Privacy Dashboard* before. Most of them stated they used it less than once a month ($N = 5$), and a few used it at least once a month ($N = 2$). Two participants stated that they had not used the Dashboard before, while 11 were unaware of the Dashboard at all. Note that these low numbers need to be treated with care and are expected, as the goal of the Dashboard is to provide an overview when the need arises. Thus, low usage counts are expected.

4.3 RQ1: Awareness of Permission States

In daily ESM awareness questionnaires, we asked participants if they were aware of the current privacy permissions of certain apps (up to five apps per questionnaire, randomly chosen). Questions were in the form of "Does app x cur-

rently have access to permission y?". Overall, 885 daily ESM awareness questionnaires were answered by participants (6.70 on average per participant), covering 4,395 questions (i.e., permission-app tuples): 2,153 (49%) questions targeted permissions currently granted and 2,242 (51%) currently denied.

For permissions that were currently granted, this was assessed correctly almost half of the time ($N = 1,052$, 49%), 455 times (21%) it was falsely believed the permission was currently not granted, and 646 times (30%) participants indicated they do not know. For permissions that were currently not granted, 760 questions were correctly answered with "No" (34%). In contrast, 685 were falsely answered with "Yes" (i.e., granted, 31%) and 797 times participants indicated they do not know (36%)[6]. Table 3 provides an overview. Cases where specific permissions were granted, but participants believed they were denied (i.e., they answered with "No"), are particularly privacy-critical. Looking at the specific permissions that we (randomly) asked for in the ESM awareness questionnaires, the following permissions were often falsely assessed: read (70 of 325) and write (51 of 223) access to the external storage; read phone state (33 of 167); and camera access (31 of 129). Table 8 in Appendix D.2 provides details on correct and false answers per permission.

4.4 RQ2: Controlling Permissions

We collected data of 2,866 updates on privacy permissions, including revoking permissions previously granted and vice versa. Data was collected automatically through scans of our study app and manually using experience sampling (ESM control questionnaire), asking participants for reasons for their permission updates (see Tables 5 and 4 for an overview). Participants mainly chose among the given answer options, while "Other" was chosen only 40 times (1122 ESM questions in total, with 1289 reasons given). Given the low number, we report these examples directly where appropriate.

4.4.1 Revoked Permissions

Of 2,866 permission updates, 1,802 were revocations (62.88%). Participants mostly revoked read ($N = 276$) and write ($N = 242$) access to their external storage, camera access ($N = 192$), location access (coarse $N = 155$, fine $N = 142$, and background $N = 56$), and permissions to record audio ($N = 136$), read phone state ($N = 125$) or contacts ($N = 90$), get accounts ($N = 83$), Bluetooth scans ($N = 74$), and query all packages ($N = 71$). Table 9 in Appendix D.3 provides an overview of the apps most affected by revokes. Interestingly, revokes also affected apps heavily used, including Instagram ($N = 11$), TikTok ($N = 27$), YouTube ($N = 18$), or Messenger ($N = 13$, Table 6). Revoked permissions (e.g., location for Instagram or TikTok) are not essential for *consuming* content.

---

[5]Note that this is in line with the importance of how Android 11+ treats these permissions, e.g. by allowing one-time permissions [10].

[6]Note that some of these "Yes" answers may stem from the auto-reset feature of Android 11/12 having revoked permissions automatically. Hence, some "Yes" answers might have been correct from the participants' point-of-view. Nonetheless, we did not observe differences between users with Android 11/12 and those with older versions (see Table 3).

Table 3: Awareness of Current Privacy Permissions: Participants' answers to daily ESM awareness questionnaires. For "granted" permissions, the correct answer is "yes", while for "denied" permissions, the correct answer is "no" (marked in bold/green). The table shows the total distribution of answers and for older (8-10) vs newer (11-12) Android versions.

Permission State	Yes		No		I don't know
Granted	**1052 (49%)** old: 54%, new: 44%		455 (21%) old: 22%, new: 21%		646 (30%) old: 24%, new: 35%
Denied	685 (31%) old: 31%, new: 30%		**760 (34%)** old: 22%, new: 32%		797 (36%) old: 24%, new: 38%

Table 4: Reasons for *Revoking* Permissions: Using Experience Sampling, we gathered the reasons for a total of 682 revokes that were conducted during the study (multiple select).

Reason for *Granting* Permissions		
I didn't need the feature.	357	
I was concerned about my privacy.	212	
I didn't think about it.	135	
I was concerned about the security of my device.	88	
Other	22	

Table 5: Reasons for *Granting* Permissions: Using Experience Sampling, we gathered the reasons for a total of 440 grants that were conducted during the study (multiple select).

Reason for *Granting* Permissions		
I wanted to enable a feature of the app.	227	
The app asked for it.	117	
I didn't think about it.	106	
Other	18	

Using the ESM control questionnaires, we acquired additional data on 682 revocation events. The ESM mainly covered events related to revoking read ($N = 120$) and write ($N = 103$) access to external storage, access to the camera ($N = 98$) and location (coarse $N = 62$ and fine $N = 47$), reading phone states ($N = 44$), recording audio ($N = 42$), or reading contacts ($N = 31$). Apps that were covered mostly include TikTok ($N = 17$), ZAFUL ($N = 12$), Twitter ($N = 12$), PayPal ($N = 12$), Pikmin Bloom ($N = 9$), and others. As reasons for their decision, participants mostly mentioned not needing the respective feature (mentioned for 352 revokes by 75 participants), privacy (212 revokes, 58 participants), and security (88 revokes, 36 participants) concerns, not having thought about it (133 revokes, 43 participants), and other (22 revokes, 11 participants) such as they did not actively choose or the app did not ask for it, or "it must have happened automatically" (one participant each, see Table 4). Participants could choose several reasons when asked for a specific app and permission.

4.4.2 Permissions Granted Later

Updates during the study included 1,064 'granted later' permissions (37.12%). Participants mostly granted permission to read ($N = 121$) or write ($N = 101$) external storage, access location (coarse $N = 101$; fine $N = 91$), record audio ($N = 136$), query all packages ($N = 74$), read phone state ($N = 59$), and camera ($N = 54$). Regarding highly used apps, grants affected,

e.g., Instagram ($N = 3$) or TikTok ($N = 21$), but also apps of the category *Tools* such as Google ($N = 25$) or the Phone ($N = 31$, see Table 6). Table 10 in Appendix D.3 provides an overview of apps with most grants.

From the ESM control questionnaires, we acquired data on 440 permission grants, mainly affecting permission to read ($N = 81$) and write ($N = 63$) external storage, access to Bluetooth ($N = 54$), camera ($N = 42$), or location (coarse $N = 38$ and fine $N = 29$), or to query all packages ($N = 30$). Apps affected included Instagram ($N = 9$), Ferrarm SIM ($N = 9$), and others. Participants mostly wanted to enable a feature of an app (mentioned for 223 grants by 70 participants), and the affected app asked for certain permission (117 grants, 50 participants), or they did not think about it (105 grants, 38 participants). For 18 grants, other reasons were mentioned (11 participants), including they did not remember giving permission, were unsure about consequences, or the app was pre-installed (one participant each, see Table 5). Note that participants could choose several reasons again.

4.4.3 Bulk Permission Updates

A total of 702 apps were affected by permission updates throughout the study (692 unique on a per-user basis, 493 unique apps overall), with updates of 2,866 permissions in total (22.39 on average across 128 participants who conducted such updates). Many scans by our study app (every two hours) comprised updates of more than one app and/or more than one permission, indicating that participants ($N = 128$) conducted updates in "bulks", that is, in short time frames. Such scans included one to 27 apps ($Mean = 1.43$, $SD = 1.57$), with 106 updates including more than one app (384 updates included only one). Per app, more than one permission was updated in most cases ($N = 487$ vs. 215 cases with single permissions updated for an app), with 1 to 19 permissions updated at once ($Mean = 4.08$, $SD = 3.44$). In total, 383 scans included updates of multiple permissions (107 scans only one).

4.4.4 User vs. System-initiated Permission Updates

A total of 66 participants (50%) were using Android 11 or above. For these, Android might have initiated some permission updates. In particular, the *auto revoke* [9] feature automatically withdraws permissions for apps that have not been used for several months. However, among the 1392 permission updates (808 revokes) we collected from participants with newer Android versions, only 32 revokes (4%) affected apps not used at all during the two weeks of study. More-

Table 6: Most Used Apps and Permission Updates: This table presents the most used apps in our dataset (left) along with corresponding permission updates (right). We sorted all *apps* based on the total overall *usage time* (sum in hours) as acquired from the initial scan. We list the first 25 apps below (N: number of installations). The *Category* is based on the Google Playstore [2], except for two side-loaded apps that we categorized accordingly. Permission updates include *revokes* and *grants*.

App	N	Category	Total Usage Time (hours)	Permission Updates: **Revokes**		Permission Updates: **Grants**
Instagram	94	Social	1980.9	11	loc_c(1), loc_f(1), media(1), cam(1), acct(1), cont_r(1), stor_r(1), phone#(1), phstate(1), audio(1), stor_w(1)	3 · loc_c(1), bt(1), bt_s(1)
Chrome	131	Communications	1581.1	1	bt_s(1)	0
TikTok	62	Social	1494.1	27	pkg(11), loc_c(2), cam(2), cal_r(2), cont_r(2), stor_r(2), audio(2), cal_w(2), stor_w(2)	21 · pkg(18), bt(2), bt_s(1)
Facebook	93	Social	1345.1	1	phone#(1)	1 · bt_s(1)
Messenger	100	Communications	1003.6	13	loc_c(1), loc_f(1), cam(1), acct(1), cal_r(1), cont_r(1), stor_r(1), phone#(1), phstate(1), sms(1), audio(1), cont_w(1), stor_w(1)	2 · bt(1), bt_s(1)
YouTube	121	Video Players & Editors	888.2	18	loc_c(2), loc_f(2), cam(2), acct(2), cont_r(2), stor_r(2), phstate(2), audio(2), stor_w(2)	0
WhatsApp	87	Communications	684.1	0		1 · bt_s(1)
Reddit	51	Social	467.0	17	loc_c(4), cam(4), stor_r(4), audio(4), stor_w(1)	5 · bt(4), bt_s(1)
YouTube Vanced	20	Video Players & Editors	427.2	11	loc_b(1), loc_c(1), loc_f(1), media(1), cam(1), acct(1), cont_r(1), stor_r(1), phstate(1), audio(1), stor_w(1)	0
Huawei Home	14	Tools	411.5	2	cal_r(2)	2 · bt(2)
Telegram	50	Communications	304.0	21	loc_b(2), loc_c(2), loc_f(2), cam(2), acct(2), cont_r(2), stor_r(2), audio(2), cont_w(2), stor_w(2), call(1)	8 · bt_s(1), bt(2), call(1), phone#(2), phstate(2)
Twitter	48	Social	280.8	18	loc_c(2), loc_f(2), cam(2), acct(2), cont_r(2), stor_r(2), phstate(2), audio(2), stor_w(2)	3 · bt(3)
Netflix	68	Entertainment	273.2	0		0
Discord	63	Communications	262.3	1	cont_r(1)	5 · bt(1), cam(1), stor_r(1), audio(1), bt_s(1)
Google	123	Tools	208.0	13	loc_b(2), act(2), bt_s(2), pkg(2), stor_r(2), stor_w(2), cam(1)	25 · loc_c(2), loc_f(2), bt(2), cam(1), acct(2), cal_r(2), call(2), cont_r(2), phstate(2), sms(2), audio(2), cal_w(2), cont_w(2)
Phone	105	Tools	150.5	0		31 · loc_c(3), bt(3), acct(3), call(3), cont_r(3), stor_r(2), phstate(2), sms(3), audio(3), cont_w(3), stor_w(2), loc_f(1)
Snapchat	35	Communications	116.4	28	phone#(4), loc_b(3), loc_c(2), loc_f(2), cam(1), acct(3), call(3), cont_r(2), stor_r(1), phstate(2), audio(1), cont_w(3), stor_w(1)	17 · bt(3), bt_s(1), loc_c(1), loc_f(1), cam(2), cont_r(1), stor_r(2), phone#(1), phstate(1), audio(1), stor_w(2)
Spotify	89	Music & Audio	112.0	3	bt_s(3)	1 · bt(1)
Clock	87	Tools	102.1	1	phstate(1)	0
Maps	78	Travel & Local	96.7	9	loc_b(1), media(1), act(1), bt_s(1), cam(1), cont_r(1), stor_r(1), audio(1), stor_w(1)	3 · loc_c(1), loc_f(1), bt(1)
Gallery	82	Photography	96.0	11	loc_c(3), loc_f(3), pkg(2), acct(1), cont_r(1), phstate(1)	5 · media(2), pkg(1), stor_r(1), stor_w(1)
Zoom	18	Business	73.7	0		0
Gmail	127	Communications	73.6	1	phstate(1)	0

✈_f access fine location; ✈_c access coarse location; ✈_b access background location; 📅_r read calendar; 📅_w write calendar; 🖼 access media location; 📷 camera; ᛒ Bluetooth; ᛒ_s Bluetooth scan; 💾_r read external storage; 💾_w write external storage; ▢ query all packages; 🎤 record audio; 👥_r read contacts; 👥_w write contacts; ≣ read phone numbers; 🔑 get accounts; ▯ read phone state; 🏃 activity recognition; 💬 read SMS; 📞 read call log;

over, we collected permission revocations among users in both groups ($N = 61$ users with permission revokes on new Android, $N = 64$ users with revokes on old Android). The total number of permissions updated is very similar: $N = 1,474$ for older Android versions and $N = 1,392$ for newer Android versions[7]. In addition, many permission updates (including

revokes) were conducted for heavily used apps (see Table 6). Thus, the majority of revoked permissions do not fall under the auto-revoke feature. Still, there might be cases in which users chose to grant camera, location, or microphone permissions for *one time* [10] only. These permissions are revoked automatically as soon as the requesting app moves into the background. Hence, such one-time permissions would only

---

[7] Note that we did not find any statistically significant differences in the number of updates (neither for total number nor number of grants or revokes)

for users with older vs newer Android versions.

occur in our dataset if users used an app during the end of/the beginning of a new two-hour timeslot. To identify such cases, we looked at permissions per user and app that were granted and revoked multiple times but found no such cases.

5 Discussion

5.1 Awareness of Privacy Permission States

Our study results indicate that people have an alarmingly low level of awareness regarding what permissions specific apps have. In particular, only 49% of granted permissions and 34% of denied permissions were assessed correctly. Past studies have shown that privacy awareness is a prerequisite for users to make meaningful decisions [29, 65, 66], for example, about whether or not an app should retain certain permission at a given point in time. Moreover, Frik et al. identified a lack of awareness regarding the availability of privacy settings, leading users to not take action according to their privacy needs [42]. The permission model of modern smartphone operating systems seems to address this already: asking for permissions in context, that is, right at the time when they are needed (cf. *runtime permissions* on newer Android versions). This helps users build better mental models of the permission space and also enables them to select only permissions that make sense to them for a particular application or feature.

We found that in many cases users think that permissions they gave are not actually given and vice versa. In 455 cases, granted permissions were falsely assessed as denied (21%), which is critical from a privacy point-of-view as apps might access personal data without users being aware of it. Moreover, there were many cases where users indicated they did not know (granted: 30%, denied: 36%). This is likely related to the sheer number of (partially unused or rarely used) installed applications on users' phones (99.42 on average for this study). Also, prior work found that it is oftentimes not clear to users which permissions are requested for the actual application vs. third-party services [33], and textual descriptions of applications oftentimes lack detailed information on permission requests [37]. Permission reminders, as standard in current OS versions and other proactive features, can mitigate this to some extent but come with the risk of overburdening users with recurring warnings. However, as shown in previous work, increasing awareness can help to motivate them and take action about their privacy [34, 45, 73], and information prompts might thus be acceptable. Information that could be relevant in such interfaces includes, but is not limited to, the type of data that is collected and stored, for how long, and with whom it is shared [62]. Other relevant information includes whether an app can access private data in the background; or how an app is rated by others [77]. Another opportunity could be to convey information on the risks rather than the resources or sensors being assessed [40].

5.2 Types of Permission Revocations

Participants changed permission states for installed apps in 2,866 cases, including 1,802 *revocations*. Reasons include a lack of need for (or lost benefit of) the respective feature, privacy, and security, in that order (see Table 4). This indicates that, while privacy and security play a major role in the process (mentioned by 94 participants, 71%), other factors do as well in a significant way. This indicates that messaging around revocation support should cater to these needs to help users make informed decisions. Looking at the details of permission revocations, three trends become apparent:

1) **Privacy-Relevance of Revoked Permissions.** Most revocations recorded in the study fall into the bucket of the top three permissions that participants want to have an eye on: location, camera, and microphone access. The sensitivity of this data was shown in previous work [18, 22, 27, 29, 57, 60, 69].

2) **Affected Apps.** Participants rarely revoked permissions for frequently used apps. This indicates that the benefit of allowing an app access to a certain permission increases through usage frequency. While YouTube and TikTok are among the most installed apps with high usage time showing relevant revocation activity, their revocations fall into what will be listed in point 3 below.

3) **Functionality-Relevance of Revoked Permissions.** Revocations are mostly related to permissions not essential for the app's core functionality. In particular, looking at the apps with the highest usage times (see Table 6), permissions might be necessary for *producing* content but not for *consumption*. For instance, access to the camera, external storage, media, or location was revoked for apps such as Instagram, TikTok, and Youtube, which still allows using these apps to *consume* content. This indicates that users consider the use case and functionality they intend to use an app for when deciding on permissions and use the opportunity to restrict permissions necessary for *producing* content if they do not intend to do so. Thus, future approaches could consider effects on core functionality [51]. This is in line with prior work indicating that end-users and developers alike consider app functionality and features when it comes to permission management [82].

Researchers have suggested a number of innovative privacy designs. The above-mentioned trends and the insights from our real-world study build a solid foundation to review existing privacy designs and assess their ability to address important aspects we identified (see below). Furthermore, they can inform future designs of mobile privacy control.

5.3 (Proactive) Mobile Privacy Control

Prior work showed users want to protect personally identifiable information on their smartphones and, thus, are open to supportive tools [23]. Privacy protection mechanisms follow different approaches in terms of proactivity, from low to

high [52] or from simple notifications (or recommendations) to full automation, where systems act on users' behalf [30]. Users prefer simple and proactive mechanisms while still staying in control as opposed to full automation [30, 52]. Moreover, permission prompts should provide explanations to increase users' confidence in their decisions [35].

Finding the right balance for proactive privacy support features on mobile devices seems, thus, essential. For instance, proactive privacy permissions (as an extension to the current runtime permission model) could a) learn over time or b) be based on rules (for example, context-based) [64]. Proactive privacy controls could also guide users through available settings [42] or notify them when in privacy-critical contexts [30] to, for example, avoid microphone access in private spaces or at custom timings [41, 67]. Alternatively, privacy controls could adapt to users' profiles and, e.g., suggest revoking certain permissions vs. entirely uninstalling certain apps [22].

Our study data shows that before and after the study, participants were interested in getting proactive support, such as being notified, especially regarding permissions related to location, camera, and microphone use. While these permissions are essential for using some applications (e.g., a microphone for using the phone), in many cases, permissions are secondary (e.g., microphone access for a messenger application supporting text-based communication). As discussed above, other examples include revoking permissions for non-active apps, as implemented since Android 11, and for non-essential vs. essential permissions for a certain app. A proactive privacy control mechanism could focus on permissions users care about from a privacy perspective but still consider functionalities essential to users based on the intended use of a certain app (e.g., consuming vs. producing content). Considering contextual information can improve recommendations for privacy settings [71, 79]. Also, permissions non-essential for the core functionality could be detected automatically (cf. the Reaper approach [33]). Communicating to users which permissions are a pre-condition for using a certain functionality can additionally help them make a decision [76]. As such, the overall decision load could still be kept rather low.

Moreover, information on permissions that are (un)desired could be crowd-sourced based on users' comments, as suggested in prior work: CHAMP analyzes users' comments to point to undesired and/or privacy-intrusive app behaviors [50]. However, the capabilities and opportunities of novel smartphones and apps keep changing, as do users' preferences. This indicates user preferences should be assessed repeatedly.

5.4 Bulk Revocations & Opportune Moments

When participants updated permissions for a specific app, independent of the (external) trigger, they often seemed to engage in updating more permissions for the given app (4.08 permissions per application on average) as well as permissions for other apps (in 106 cases) in short time frames. This is interesting for two reasons: First, current privacy interfaces on mobile devices such as Android's *Permission Manager* and *Privacy Dashboard* already foster bulk permission updates by displaying other apps with the same permission or other permissions for the same app. With knowledge of applications for which users jointly change permissions, permission management interfaces could proactively suggest groups of apps for which a particular permission could be changed. Second, this point in time represents an opportune moment in which users are willing and motivated to engage in a privacy/security activity. Due to the two-hour time window of our study, we were not able to explore those opportune moments in more detail, but future work could look at phone usage patterns, users' current mood or necessity of the current privacy decision [31]. Leveraging this information could further support users in maintaining correct permission states for them.

6   Conclusion

We presented an in-depth investigation of users' awareness of and control over privacy permissions on Android. In a two-week field study with 132 Android users, we collected initial permission states of installed applications as well as updates of permission states throughout the study and experience sampling data. We found that participants mostly revoked access to sensors they consider sensitive (such as microphone or camera), but only if this would not affect an application's core functionality, assuming the app is frequently used. Moreover, participants often conducted such permission updates in bulk. This work provides a better understanding of users' current use of available privacy control mechanisms and serves as a basis to enhance (proactive) mobile privacy control.

References

[1] Android 12. https://www.android.com/android-12/, 2022. last accessed: 2023-02-15.

[2] Choose a category and tags for your app or game. https://support.google.com/googleplay/android-developer/answer/9859673?hl=en, 2022. last accessed: 2023-02-15.

[3] Firebase App Distribution. https://firebase.google.com/docs/app-distribution, 2022. last accessed: 2023-02-15.

[4] Firebase Crashlytics. Track, prioritize, and fix crashes faster. https://firebase.google.com/products/crashlytics, 2022. last accessed: 2023-02-15.

[5] Firebase Realtime Database. Store and sync data in real time. https://firebase.google.com/products/realtime-database, 2022. last accessed: 2023-02-15.

[6] How do I balance my sample within demographics? https://researcher-help.prolific.co/hc/

`en-gb/articles/360009221213`, 2022. last accessed: 2023-02-15.

[7] MPAndroidChart. `https://github.com/PhilJay/MPAndroidChart`, 2022. last accessed: 2023-02-15.

[8] Permissions on Android. `https://developer.android.com/guide/topics/permissions/overview`, 2022. last accessed: 2023-02-15.

[9] Permissions updates in Android 11. `https://developer.android.com/about/versions/11/privacy/permissions`, 2022. last accessed: 2023-02-15.

[10] Permissions updates in Android 11. One-time permissions. `https://developer.android.com/about/versions/11/privacy/permissions#one-time`, 2022. last accessed: 2023-02-15.

[11] Photo picker. `https://developer.android.com/training/data-storage/shared/photopicker`, 2022. last accessed: 2023-02-15.

[12] Prolific. A higher standard of online research. `https://prolific.co/`, 2022. last accessed: 2023-02-15.

[13] Request app permissions. `https://developer.android.com/training/permissions/requesting`, 2022. last accessed: 2023-02-15.

[14] SurveyKit: Create beautiful surveys on Android. `https://github.com/quickbirdstudios/SurveyKit`, 2022. last accessed: 2023-02-15.

[15] What's new in Kotlin 1.6.20. `https://kotlinlang.org/docs/whatsnew1620.html`, 2022. last accessed: 2023-02-15.

[16] Android Developers - Documentation. PackageManager. `https://developer.android.com/reference/android/content/pm/PackageManager`, 2023. last accessed: 2023-05-15.

[17] Android Developers - Documentation. UsageStatsManager. `https://developer.android.com/reference/android/app/usage/UsageStatsManager`, 2023. last accessed: 2023-05-15.

[18] Noura Abdi, Kopo M. Ramokapane, and Jose M. Such. More than Smart Speakers: Security and Privacy Perceptions of Smart Home Personal Assistants. In *Proceedings of the Symposium on Usable Privacy and Security*, SOUPS '19, pages 1–16, Berkeley, CA, USA, 2019. USENIX Association.

[19] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. Privacy and human behavior in the age of information. *Science*, 347(6221):509–514, 2015.

[20] Imtiaz Ahmad, Rosta Farzan, Apu Kapadia, and Adam J. Lee. Tangible Privacy: Towards User-Centric Sensor Designs for Bystander Privacy. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2), October 2020.

[21] Hazim Almuhimedi, Florian Schaub, Norman Sadeh, Idris Adjerid, Alessandro Acquisti, Joshua Gluck, Lorrie Faith Cranor, and Yuvraj Agarwal. Your Location Has Been Shared 5,398 Times! A Field Study on Mobile App Privacy Nudging. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 787–796, New York, NY, USA, 2015. Association for Computing Machinery.

[22] Ashwaq Alsoubai, Reza Ghaiumy Anaraky, Yao Li, Xinru Page, Bart Knijnenburg, and Pamela J. Wisniewski. Permission vs. App Limiters: Profiling Smartphone Users to Understand Differing Strategies for Mobile Privacy Management. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery.

[23] Evita Bakopoulou, Anastasia Shuba, and Athina Markopoulou. Exposures Exposed: A Measurement and User Study to Assess Mobile Data Privacy in Context, 2020.

[24] David G. Balash, Xiaoyuan Wu, Miles Grant, Irwin Reyes, and Adam J. Aviv. Security and privacy perceptions of Third-Party application access for google accounts. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 3397–3414, Boston, MA, August 2022. USENIX Association.

[25] Bram Bonné, Sai Teja Peddinti, Igor Bilogrevic, and Nina Taft. Exploring decision making with Android's runtime permission dialogs using in-context surveys. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*, pages 195–210, Santa Clara, CA, July 2017. USENIX Association.

[26] Weicheng Cao, Chunqiu Xia, Sai Teja Peddinti, David Lie, Nina Taft, and Lisa M. Austin. A Large Scale Study of User Behavior, Expectations and Engagement with Android Permissions. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 803–820. USENIX Association, August 2021.

[27] George Chalhoub, Martin J Kraemer, Norbert Nthala, and Ivan Flechais. "It Did Not Give Me an Option to Decline": A Longitudinal Analysis of the User Experience of Security and Privacy in Smart Home Products.

In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[28] Richard Chow, Serge Egelman, Raghudeep Kannavara, Hosub Lee, Suyash Misra, and Edward Wang. HCI in Business: A Collaboration with Academia in IoT Privacy. In Fiona Fui-Hoon Nah and Chuan-Hoo Tan, editors, *HCI in Business*, pages 679–687, Cham, 2015. Springer International Publishing.

[29] Camille Cobb, Sruti Bhagavatula, Kalil Anderson Garrett, Alison Hoffman, Varun Rao, and Lujo Bauer. "I would have to evaluate their objections": Privacy tensions between smart home device owners and incidental users. *Proceedings on Privacy Enhancing Technologies*, 4:54–75, 2021.

[30] Jessica Colnago, Yuanyuan Feng, Tharangini Palanivel, Sarah Pearman, Megan Ung, Alessandro Acquisti, Lorrie Faith Cranor, and Norman Sadeh. Informing the Design of a Personalized Privacy Assistant for the Internet of Things. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA, 2020. Association for Computing Machinery.

[31] Jessica Colnago and Hélio Guardia. How to Inform Privacy Agents on Preferred Level of User Control? In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, UbiComp '16, page 1542–1547, New York, NY, USA, 2016. Association for Computing Machinery.

[32] Lorrie Faith Cranor. Necessary but not sufficient: Standardized mechanisms for privacy notice and choice. *J. on Telecomm. & High Tech. L.*, 10:273, 2012.

[33] Michalis Diamantaris, Elias P. Papadopoulos, Evangelos P. Markatos, Sotiris Ioannidis, and Jason Polakis. REAPER: Real-Time App Analysis for Augmenting the Android Permission System. In *Proceedings of the Ninth ACM Conference on Data and Application Security and Privacy*, CODASPY '19, page 37–48, New York, NY, USA, 2019. Association for Computing Machinery.

[34] Serge Egelman, Sakshi Jain, Rebecca S. Portnoff, Kerwell Liao, Sunny Consolvo, and David Wagner. Are You Ready to Lock? In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, CCS '14, page 750–761, New York, NY, USA, 2014. Association for Computing Machinery.

[35] Yusra Elbitar, Michael Schilling, Trung Tin Nguyen, Michael Backes, and Sven Bugiel. Explanation beats context: The effect of timing & rationales on users'

runtime permission decisions. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 785–802. USENIX Association, August 2021.

[36] Pardis Emami-Naeini, Sruti Bhagavatula, Hana Habib, Martin Degeling, Lujo Bauer, Lorrie Cranor, and Norman Sadeh. Privacy Expectations and Preferences in an IoT World. In *Proceedings of the Symposium on Usable Privacy and Security*, SOUPS '17, pages 399–412, Berkeley, CA, USA, 2017. USENIX Association.

[37] Johannes Feichtner and Stefan Gruber. Understanding Privacy Awareness in Android App Descriptions Using Deep Learning. In *Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy*, CODASPY '20, page 203–214, New York, NY, USA, 2020. Association for Computing Machinery.

[38] Adrienne Porter Felt, Erika Chin, Steve Hanna, Dawn Song, and David Wagner. Android Permissions Demystified. In *Proceedings of the 18th ACM Conference on Computer and Communications Security*, CCS '11, page 627–638, New York, NY, USA, 2011. Association for Computing Machinery.

[39] Adrienne Porter Felt, Serge Egelman, and David Wagner. I've Got 99 Problems, but Vibration Ain't One: A Survey of Smartphone Users' Concerns. In *Proceedings of the Second ACM Workshop on Security and Privacy in Smartphones and Mobile Devices*, SPSM '12, page 33–44, New York, NY, USA, 2012. Association for Computing Machinery.

[40] Adrienne Porter Felt, Elizabeth Ha, Serge Egelman, Ariel Haney, Erika Chin, and David Wagner. Android Permissions: User Attention, Comprehension, and Behavior. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, SOUPS '12, New York, NY, USA, 2012. Association for Computing Machinery.

[41] Yuanyuan Feng, Yaxing Yao, and Norman Sadeh. A Design Space for Privacy Choices: Towards Meaningful Privacy Control in the Internet of Things. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[42] Alisa Frik, Juliann Kim, Joshua Rafael Sanchez, and Joanne Ma. Users' Expectations About and Use of Smartphone Privacy and Security Settings. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery.

[43] Radhika Garg and Christopher Moreno. Understanding Motivators, Constraints, and Practices of Sharing Internet of Things. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 3(2), June 2019.

[44] Nina Gerber, Paul Gerber, and Melanie Volkamer. Explaining the privacy paradox: A systematic review of literature investigating privacy attitude and behavior. *Computers & Security*, 77:226–261, 2018.

[45] Nina Gerber, Benjamin Reinheimer, and Melanie Volkamer. Investigating People's Privacy Risk Perception. *Proceedings on privacy enhancing technologies*, 2019(3):267–288, 2019.

[46] Hana Habib and Lorrie Faith Cranor. Evaluating the Usability of Privacy Choice Mechanisms. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 273–289, Boston, MA, August 2022. USENIX Association.

[47] Weijia He, Maximilian Golla, Roshni Padhi, Jordan Ofek, Markus Dürmuth, Earlence Fernandes, and Blase Ur. Rethinking Access Control and Authentication for the Home Internet of Things (IoT). In *27th USENIX Security Symposium (USENIX Security 18)*, pages 255–272, Baltimore, MD, August 2018. USENIX Association.

[48] Yangyang He. Recommending Privacy Settings for IoT. In *Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion*, IUI '19, page 157–158, New York, NY, USA, 2019. Association for Computing Machinery.

[49] Franziska Herbert, Gina Maria Schmidbauer-Wolf, and Christian Reuter. Who Should Get My Private Data in Which Case? Evidence in the Wild. In *Mensch Und Computer 2021*, MuC '21, page 281–293, New York, NY, USA, 2021. Association for Computing Machinery.

[50] Yangyu Hu, Haoyu Wang, Tiantong Ji, Xusheng Xiao, Xiapu Luo, Peng Gao, and Yao Guo. CHAMP: Characterizing Undesired App Behaviors from User Comments Based on Market Policies. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, pages 933–945, 2021.

[51] Qatrunnada Ismail, Tousif Ahmed, Kelly Caine, Apu Kapadia, and Michael K Reiter. To Permit or Not to Permit, That is the Usability Question: Crowdsourcing Mobile Apps' Privacy Permission Settings. *Proceedings on Privacy Enhancing Technologies*, 2017(4):119–137, 2017.

[52] Haojian Jin, Boyuan Guo, Rituparna Roychoudhury, Yaxing Yao, Swarun Kumar, Yuvraj Agarwal, and Jason I. Hong. Exploring the Needs of Users for Supporting Privacy-Protective Behaviors in Smart Homes. In *CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery.

[53] Patrick Gage Kelley, Sunny Consolvo, Lorrie Faith Cranor, Jaeyeon Jung, Norman Sadeh, and David Wetherall. A Conundrum of Permissions: Installing Applications on an Android Smartphone. In Jim Blyth, Sven Dietrich, and L. Jean Camp, editors, *Financial Cryptography and Data Security*, pages 68–79, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[54] Patrick Gage Kelley, Lorrie Faith Cranor, and Norman Sadeh. Privacy as Part of the App Decision-Making Process. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, page 3393–3402, New York, NY, USA, 2013. Association for Computing Machinery.

[55] Jennifer King, Airi Lampinen, and Alex Smolen. Privacy: Is There an App for That? In *Proceedings of the Seventh Symposium on Usable Privacy and Security*, SOUPS '11, New York, NY, USA, 2011. Association for Computing Machinery.

[56] Agnieszka Kitkowska, Mark Warner, Yefim Shulman, Erik Wästlund, and Leonardo A. Martucci. Enhancing Privacy through the Visual Design of Privacy Notices: Exploring the Interplay of Curiosity, Control and Affect. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*, pages 437–456, Berkeley, CA, USA, August 2020. USENIX Association.

[57] Predrag Klasnja, Sunny Consolvo, Tanzeem Choudhury, Richard Beckwith, and Jeffrey Hightower. Exploring Privacy Concerns about Personal Sensing. In Hideyuki Tokuda, Michael Beigl, Adrian Friday, A. J. Bernheim Brush, and Yoshito Tobe, editors, *Pervasive Computing*, pages 176–183, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[58] Konrad Kollnig, Anastasia Shuba, Reuben Binns, Max Van Kleek, and Nigel Shadbolt. Are iphones really better for privacy? a comparative study of ios and android apps. *Proceedings on Privacy Enhancing Technologies*, 2022(2):6–24, March 2022.

[59] Scott Lederer, Jennifer Mankoff, and Anind K. Dey. Who Wants to Know What When? Privacy Preference Determinants in Ubiquitous Computing. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '03, page 724–725, New York, NY, USA, 2003. Association for Computing Machinery.

[60] H. Lee and A. Kobsa. Understanding user privacy in Internet of Things environments. In *2016 IEEE 3rd World Forum on Internet of Things (WF-IoT)*, pages 407–412, New York, NY, USA, 2016. IEEE.

[61] H. Lee and A. Kobsa. Privacy preference modeling and prediction in a simulated campuswide IoT environment.

In *2017 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 276–285, New York, NY, USA, 2017. IEEE.

[62] Pedro Giovanni Leon, Blase Ur, Yang Wang, Manya Sleeper, Rebecca Balebako, Richard Shay, Lujo Bauer, Mihai Christodorescu, and Lorrie Faith Cranor. What Matters to Users? Factors That Affect Users' Willingness to Share Information with Online Advertisers. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, SOUPS '13, New York, NY, USA, 2013. Association for Computing Machinery.

[63] Bin Liu, Mads Schaarup Andersen, Florian Schaub, Hazim Almuhimedi, Shikun Zhang, Norman Sadeh, Alessandro Acquisti, and Yuvraj Agarwal. Follow my recommendations: A personalized privacy assistant for mobile app permissions. In *Proceedings of the Twelfth USENIX Conference on Usable Privacy and Security*, SOUPS '16, page 27–41, USA, 2016. USENIX Association.

[64] Nathan Malkin, David Wagner, and Serge Egelman. Runtime Permissions for Privacy in Proactive Intelligent Assistants. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 633–651, Boston, MA, August 2022. USENIX Association.

[65] Karola Marky, Sarah Prange, Florian Krell, Max Mühlhäuser, and Florian Alt. "You Just Can't Know about Everything": Privacy Perceptions of Smart Home Visitors. In *19th International Conference on Mobile and Ubiquitous Multimedia*, page 83–95, New York, NY, USA, 2020. Association for Computing Machinery.

[66] Karola Marky, Alexandra Voit, Alina Stöver, Kai Kunze, Svenja Schröder, and Max Mühlhäuser. "I Don't Know How to Protect Myself": Understanding Privacy Perceptions Resulting from the Presence of Bystanders in Smart Environments. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*, NordiCHI '20, New York, NY, USA, 2020. Association for Computing Machinery.

[67] Vikram Mehta, Daniel Gooch, Arosha Bandara, Blaine Price, and Bashar Nuseibeh. Privacy Care: A Tangible Interaction Framework for Privacy Management. *ACM Trans. Internet Technol.*, 21(1), February 2021.

[68] Moses Namara, Reza Ghaiumy Anaraky, Pamela Wisniewski, Xinru Page, and Bart P. Knijnenburg. Examining Power Use and the Privacy Paradox between Intention vs. Actual Use of Mobile Applications. In *European Symposium on Usable Security 2021*, EuroUSEC '21, page 223–235, New York, NY, USA, 2021. Association for Computing Machinery.

[69] David H. Nguyen, Alfred Kobsa, and Gillian R. Hayes. An Empirical Investigation of Concerns of Everyday Tracking and Recording Technologies. In *Proceedings of the International Conference on Ubiquitous Computing*, UbiComp '08, page 182–191, New York, NY, USA, 2008. Association for Computing Machinery.

[70] Claudia Oellers and Eva Wegner. Does germany need a (new) research ethics for the social sciences? *German Council for Social and Economic Data (RatSWD) Working Paper Series*, 2009.

[71] Katarzyna Olejnik, Italo Dacosta, Joana Soares Machado, Kévin Huguenin, Mohammad Emtiyaz Khan, and Jean-Pierre Hubaux. Smarper: Context-aware and automatic runtime-permissions for mobile devices. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 1058–1076, 2017.

[72] Stefan Palan and Christian Schitter. Prolific.ac — A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27, 2018.

[73] Sarah Prange, Niklas Thiem, Michael Fröhlich, and Florian Alt. "Secure Settings Are Quick and Easy!" – Motivating End-Users to Choose Secure Smart Home Configurations. In *Proceedings of the 2022 International Conference on Advanced Visual Interfaces*, AVI 2022, New York, NY, USA, 2022. Association for Computing Machinery.

[74] John Rothchild. Against Notice and Choice: the Manifest Failure of the Proceduralist Paradigm to Protect Privacy Online (or Anywhere Else). *Cleveland State Law Review*, 2018.

[75] Paul M Schwartz and Daniel Solove. Notice and choice: Implications for digital marketing to youth. In *The Second NPLAN/BMSG Meeting on Digital Media and Marketing to Children*, pages 1–6, 2009.

[76] William Seymour, Mark Cote, and Jose Such. Legal obligation and ethical best practice: Towards meaningful verbal consent for voice assistants. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery.

[77] Bingyu Shen, Lili Wei, Chengcheng Xiang, Yudong Wu, Mingyao Shen, Yuanyuan Zhou, and Xinxin Jin. Can systems explain permissions better? understanding users' misperceptions under smartphone runtime permission model. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 751–768. USENIX Association, August 2021.

[78] Robert H Sloan and Richard Warner. Beyond notice and choice: Privacy, norms, and consent. *J. High Tech. L.*, 14:370, 2014.

[79] Daniel Smullen and Yuanyuan Feng. The best of both worlds: Mitigating trade-offs between accuracy and user burden in capturing mobile app privacy preferences. *Proc. Priv. Enhancing Technol.*, 2020(1):195–215, 2020.

[80] Alina Stöver, Sara Hahn, Felix Kretschmer, and Nina Gerber. Investigating how users imagine their personal privacy assistant. *Proc. Priv. Enhancing Technol*, 2:384–402, 2023.

[81] Madiha Tabassum, Tomasz Kosiński, and Heather Richter Lipford. "I don't own the data": End User Perceptions of Smart Home Device Data Practices and Risks. In *Proceedings of the Fifteenth USENIX Conference on Usable Privacy and Security*, SOUPS'19, page 435–450, Berkeley, CA, USA, 2019. USENIX Association.

[82] Mohammad Tahaei, Ruba Abu-Salma, and Awais Rashid. Stuck in the permissions with you: Developer & end-user perspectives on app permissions & their privacy ramifications. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery.

[83] Niels van Berkel, Denzil Ferreira, and Vassilis Kostakos. The Experience Sampling Method on Mobile Devices. *ACM Comput. Surv.*, 50(6), dec 2017.

[84] Niels van Berkel and Vassilis Kostakos. *Recommendations for Conducting Longitudinal Experience Sampling Studies*, pages 59–78. Springer International Publishing, Cham, 2021.

[85] Primal Wijesekera, Arjun Baokar, Ashkan Hosseini, Serge Egelman, David Wagner, and Konstantin Beznosov. Android permissions remystified: A field study on contextual integrity. In *Proceedings of the 24th USENIX Conference on Security Symposium*, SEC'15, page 499–514, USA, 2015. USENIX Association.

[86] Primal Wijesekera, Arjun Baokar, Lynn Tsai, Joel Reardon, Serge Egelman, David Wagner, and Konstantin Beznosov. The feasibility of dynamically granted permissions: Aligning mobile privacy with user preferences. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 1077–1093, 2017.

[87] Primal Wijesekera, Arjun Baokar, Lynn Tsai, Joel Reardon, Serge Egelman, David Wagner, and Konstantin Beznosov. Dynamically regulating mobile application permissions. *IEEE Security & Privacy*, 16(1):64–71, 2018.

[88] Primal Wijesekera, Joel Reardon, Irwin Reyes, Lynn Tsai, Jung-Wei Chen, Nathan Good, David Wagner, Konstantin Beznosov, and Serge Egelman. Contextualizing privacy decisions for better prediction (and protection). In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–13, New York, NY, USA, 2018. Association for Computing Machinery.

[89] Yaxing Yao, Justin Reed Basdeo, Oriana Rosata Mcdonough, and Yang Wang. Privacy Perceptions and Designs of Bystanders in Smart Homes. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–24, November 2019.

[90] Noé Zufferey, Kavous Salehzadeh Niksirat, Mathias Humbert, and Kévin Huguenin. "revoked just now!" users' behaviors toward fitness-data sharing with third-party applications. *Proceedings on Privacy Enhancing Technologies*, 2023(1):21, 2023.

A   Project Material

To access the anonymized dataset and the study application, please contact the authors.

B   Android Privacy Interfaces
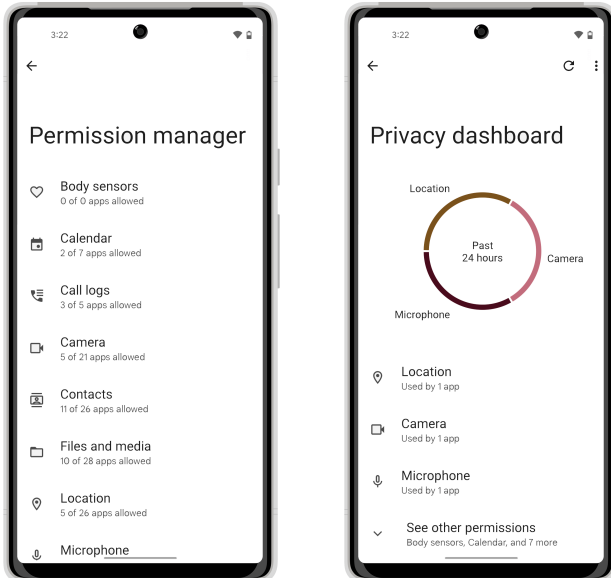
Figure 3: Android Privacy Interfaces: The *Permission Manager* (left) lists permission types along with apps that currently do or do not have access to these. The *Privacy Dashboard* (right, Android 12 and above) provides a more detailed overview of which applications currently have access to which sensors, along with means to grant or revoke this access [1].

C   Study Questionnaires

C.1   Privacy Perceptions (Initial & Final)

- For which categories of information do you want to be alerted, if an app requests access. 1/3
    - Location
    - Physical Activity
    - Body Sensors
    - Bluetooth
    - Installed apps
    - None of these

- For which categories of information do you want to be alerted, if an app requests access. 2/3
    - Other users on the smartphone
    - External File Storage
    - Calendar
    - Call history
    - Contacts
    - None of these

- For which categories of information do you want to be alerted, if an app requests access. 3/3
    - Phone Numbers
    - SMS
    - Microphone
    - Camera
    - None of these

C.2   Midterm Questionnaire

- Do you know that you can revoke permissions you previously granted to apps? (yes/no)
    - *if yes* Have you revoked a permission before? (yes/no)
        * *if yes* Why did you revoke that permission? multiple choice
            · I didn't need the feature anymore.
            · I was concerned for my privacy.
            · I was concerned for the security of my device.
            · I didn't think about it.
            · None of these.

- *For Android 12 only:* Have you used the Privacy Dashboard before? (yes/no)
    - *if yes* How often do your use the Privacy Dashboard? single choice
        * At least once a week
        * At least once a month
        * Less than once a month

- *Android versions <12:* Have you looked at the [Permission Manager name of installed Android version] before?
    - *if yes* How often do you look at the [Permission Manager name of installed Android version]? single choice
        * At least once a week
        * At least once a month
        * Less than once a month

- Please look at the [Permission Manager name of installed Android version] after our questions. You can find it under [Path].

C.3   Experience Sampling Questionnaires

C.3.1   Questions Upon Permission Change (ESM control questionnaire)

Granted Permission: Why did you allow [AppName] to [Android description for permission]? multiple choice
- I wanted to enable a feature of the app.
- The app asked for it.
- I didn't think about it.
- Other
    - Please briefly explain your decision for [AppName]. (free text entry)

Revoked Permission: Why did you forbid [AppName] to [Android description for permission]? multiple choice
- I didn't need the feature
- I was concerned for my privacy
- I was concerned for the security of my device
- I didn't think about it
- Other
    - Please briefly explain your decision for [AppName]. (free text entry)

C.3.2   Daily Questions (ESM awareness questionnaire)

- Is [App name] currently allowed to [Android description for permission]? (yes/no/I don't know)

D Study Results

D.1 Detailed Demographics

Table 7: Demographic Overview of Participant Sample: age, gender, nationality, employment, and educational level. We sampled participants by residency (not nationality), to ensure consistency in app stores. As a result, our sample contains a few participants with Asian and South American nationalities.

Age	18–29	98
	30–39	27
	40–49	5
	50–54	2
Gender	Woman (including Trans Female/Trans Woman)	65
	Man (including Trans Male/Trans Man)	63
	Non-binary (would like to give more detail)	4
Nationality	Poland	40
	Portugal	26
	Italy	22
	Greece	14
	Spain	6
	Czech Republic	4
	United Kingdom	3
	other (Europe)	12
	other (Asia)	2
	other (South America)	2
	other (North America)	1

Employment	Full-Time	38
	Unemployed (and job seeking)	32
	Other	27
	Part-Time	26
	Not in paid work (e.g. homemaker, retired or disabled)	5
	Due to start a new job within the next month	3
	Data expired	1
Education	High school diploma/A-levels	52
	Undergraduate degree (BA/BSc/other)	43
	Graduate degree (MA/MSc/MPhil/other	23
	Technical/community college	8
	Secondary education (e.g. GED/GCSE)	4
	Doctorate degree (PhD/other)	2
	Don't know / not applicable	1

D.2 RQ1: Awareness of Current Privacy Permission States

Table 8: Participants' answers to daily random questions on current permissions states. Per permission, we list the number of *correct* and incorrect answers and how often this permission and state occurred in daily random questions (in ESM awareness questionnaires) in our dataset.

Permission Name	Permission State: **Granted**			Permission State: **Denied**		
	# questions	**correct** ("yes")	**incorrect** ("no")	# questions	**correct** ("no")	**incorrect** ("yes")
write calendar	27	12	10	12	9	
read call log	46	30		6	5	3
record audio	78	57	10	159	89	28
write contacts	52	27	10	40	21	7
camera	129	75	31	224	113	73
access media location	61	35	16	26	13	8
read contacts	122	73	26	137	63	37
read external storage	325	159	70	448	138	135
access coarse location	146	95	23	162	47	60
access fine location	142	95	21	156	45	52
read phone numbers	7	7		18	5	11
read phone state	167	89	33	182	46	62
Bluetooth scan	15	6	4	38	9	2
read calendar	24	14	3	26	6	15
write external storage	223	89	51	448	100	88
get accounts	97	49	14	138	29	65
access background location	18	14	32	6	12	
query all packages	175	29	18	134	14	25
activity recognition	16	5		1	6	1
read SMS	27	12	8	7	4	3
body sensors	10	8	1	1		1

D.3 RQ2: Controlling Permissions

Table 9: Controlling Permissions: Overview of *revoked* permission updates per app throughout the study, with number of installations and total usage time. Applications that were potentially preinstalled are marked in bold, and applications that are among the most used apps (see Table 6) are marked with *. Note that only apps with at least 16 updates throughout the study are listed.

App	Number of Installs	Total Usage Time (hours)	Permission Updates: Revokes
Mi Video	51	2.35	31 ▮ Bluetooth scan (31)
Snapchat*	35	116.41	28 ▮ read phone numbers (4), access background location (3), get accounts (3), read call log (3), write contacts (3), access coarse location (2), access fine location (2), read contacts (2), read phone state (2), camera (1), read external storage (1), record audio (1), write external storage (1)
TikTok*	62	1494.08	27 ▮ query all packages (11), access coarse location (2), camera (2), read calendar (2), read contacts (2), read external storage (2), record audio (2), write calendar (2), write external storage (2)
HMS Core	18	0.51	26 ▮ access background location (2), access coarse location (2), access fine location (2), activity recognition (2), Bluetooth scan (2), camera (2), get accounts (2), query all packages (2), read calendar (2), read external storage (2), write calendar (2), write external storage (2), read contacts (1), read SMS (1)
PayPal	65	6.00	24 ▮ access coarse location (3), access fine location (3), read contacts (3), read external storage (3), read phone state (3), write contacts (3), write external storage (3)
Teams	49	16.76	21 ▮ query all packages (2), access background location (2), access coarse location (2), access fine location (2), camera (2), get accounts (2), read contacts (2), read external storage (2), write contacts (2), write external storage (2), record audio (1)
Telegram*	50	303.96	21 ▮ access background location (2), access coarse location (2), access fine location (2), camera (2), get accounts (2), read contacts (2), read external storage (2), record audio (2), write contacts (2), write external storage (2), read call log (1)
Vinted	28	42.22	18 ▮ access coarse location (3), access fine location (3), camera (3), read contacts (3), read external storage (3), write external storage (3)
Mi Browser	14	1.36	18 ▮ camera (4), record audio (4), access coarse location (3), access fine location (3), read external storage (1), write external storage (1), query all packages (1), read phone state (1)
Twitter*	48	280.76	18 ▮ access coarse location (2), access fine location (2), camera (2), get accounts (2), read contacts (2), read external storage (2), read phone state (2), record audio (2), write external storage (2)
Youtube*	121	888.21	18 ▮ access coarse location (2), access fine location (2), camera (2), get accounts (2), read contacts (2), read external storage (2), read phone state (2), record audio (2), write external storage (2)
Reddit*	51	467.04	17 ▮ access coarse location (4), camera (4), read external storage (4), record audio (4), write external storage (1)
Google Pay	39	0.16	16 ▮ access coarse location (2), access fine location (2), camera (2), query all packages (2), read contacts (2), read external storage (2), read phone state (2), write external storage (2)

Table 10: Controlling Permissions: Overview of *granted* permission updates per app throughout the study, with number of installations and average usage time. Applications that were potentially preinstalled are marked in bold, and applications that are among the most used apps (see Table 6) are marked with *. Only apps with at least 16 updates throughout the study are listed.

App	Number of Installs	Total Usage Time (hours)	Permission Updates: Grants
Phone*	105	150.47	31 ▯ access coarse location (3), Bluetooth (3), get accounts (3), read call log (3), read contacts (3), read SMS (3), record audio (3), write contacts (3), read external storage (2), read phone state (2), write external storage (2), access fine location (1)
Google*	123	207.93	25 ▯ access coarse location (2), access fine location (2), Bluetooth (2), get accounts (2), read calendar (2), read call log (2), read contacts (2), read phone state (2), read SMS (2), record audio (2), write calendar (2), write contacts (2), camera (1)
Bluetooth	37	0.17	25 ▯ access coarse location (2), Bluetooth (2), get accounts (2), read call log (2), read contacts (2), read external storage (2), read SMS (2), write contacts (2), write external storage (2), access fine location (1), access media location (1), Bluetooth scan (1), query all packages (1), read phone numbers (1), record audio (1), read phone state (1)
TikTok*	62	1494.08	21 ▯ query all packages (18), Bluetooth (2), Bluetooth scan (1)
Galaxy Store	14	1.44	18 ▯ Bluetooth (4), read phone state (4), get accounts (3), read external storage (3), write external storage (3), query all packages (1)
Snapchat*	35	116.41	17 ▯ Bluetooth (3), camera (2), read external storage (2), record audio (2), write external storage (2), Bluetooth scan (1), access coarse location (1), access fine location (1), read contacts (1), read phone numbers (1), read phone state (1)

Threat modeling state of practice in Dutch organizations

Stef Verreydt
DistriNet, KU Leuven
3001 Leuven, Belgium

Koen Yskout
DistriNet, KU Leuven
3001 Leuven, Belgium

Laurens Sion
DistriNet, KU Leuven
3001 Leuven, Belgium

Wouter Joosen
DistriNet, KU Leuven
3001 Leuven, Belgium

Abstract

Threat modeling is a key technique to apply a *security by design* mindset, allowing the systematic identification of security and privacy threats based on design-level abstractions of a system. Despite threat modeling being a best practice, there are few studies analyzing its application in practice. This paper investigates the state of practice on threat modeling in large Dutch organizations through semi-structured interviews.

Compared to related work, which mainly addresses the execution of threat modeling activities, our findings reveal multiple human and organizational factors which significantly impact the embedding of threat modeling within organizations. First, while threat modeling is appreciated for its ability to uncover threats, it is also recognized as an important activity for raising security awareness among developers. Second, leveraging developers' intrinsic motivation is considered more important than enforcing threat modeling as a compliance requirement. Third, organizations face numerous challenges related to threat modeling, such as managing the scope, obtaining relevant architectural documentation, scaling, and systematically following up on the results. Organizations can use these findings to assess their current threat modeling activities, and help inform decisions to start, extend, or reorient them. Furthermore, threat modeling facilitators and researchers may base future efforts on the challenges identified in this study.

1 Introduction

Many security-enhancing activities can be performed during software development, ranging from training and the secu-

rity requirements specification over source code analysis to pentesting and incident response handling [14]. One of these activities, and the focus of this research, is *threat modeling*.

Threat modeling is widely promoted as a best practice for secure software development. For example, it plays a prominent role in Microsoft's Security Development Lifecycle [14], OWASP's Software Assurance Maturity Model [19], NIST's Secure Software Development Framework [16], and others. Moreover, insecure design, for which threat modeling is considered a key mitigation strategy, appears in the fourth place in the most recent (2021) edition of the OWASP Top 10 [17].

In the words of the 'Threat Modeling Manifesto' [3], and in alignment with the 'four questions' framework of Shostack [24], "*threat modeling is analyzing representations of a system to highlight concerns about security and privacy characteristics. [This involves] four key questions: 1) What are we working on? 2) What can go wrong? 3) What are we going to do about it? and 4) Did we do a good enough job?*". Numerous threat modeling approaches exist (e.g., STRIDE [24], PASTA [35], CVSS [12], attack trees [22]), as well as several supporting tools (e.g., Microsoft Threat Modeling Tool [15], IriusRisk [10], pytm [32], OWASP Threat Dragon [18]) to (partially) automate threat modeling analyses.

Current empirical research (Section 5) mostly focuses on how threat modeling is applied by practitioners, identifying best practices and challenges related to specific threat modeling methodologies, tools, and application domains. Few studies, however, investigate scheduling, stakeholder involvement, frequency, organization introduction, etc, yet such non-technical aspects affect the overall effectiveness of threat modeling. Through semi-structured interviews with practitioners from large Dutch organizations in critical sectors that are part of the target audience of the Dutch National Cyber Security Center (NCSC, the sponsor of this research), the goal of this research is therefore to provide qualitative insights into the state of practice on threat modeling, paying particular attention to the non-technical aspects of threat modeling. The results of this study can be used by other organizations to assess their current practices, and help inform decisions to start, extend,

or reorient existing threat modeling programs.

The remainder of this paper is structured as follows. Section 2 explains the research methodology. Section 3 answers the research questions based on observations from the interviews. Section 4 discusses the implications of this study's results and the limitations of this study. Section 5 provides an overview of related work and relates our observations to those of similar studies. Finally, Section 6 concludes the paper.

2 Methodology

2.1 Goal and scope

The insights in this paper originate from a set of interviews with practitioners from large, Dutch organizations, conducted between August 2022 and February 2023. The sponsor of this research (NCSC) is a government organization that provides security advice to large organizations in critical sectors, and the interviewees are employees of those large organizations. We focus primarily on organizations that have in-house software development teams, but also include organizations without such teams yet focusing on Information Technology (IT) and Operational Technology (OT) infrastructure, as well as one organization that has an advisory role.

Our assessment of the state of practice addresses four broad research questions: **RQ1**: How is threat modeling embedded in the organization? **RQ2**: Which organizational roles are involved in threat modeling activities? **RQ3**: How is threat modeling performed within the organization? **RQ4**: What are the experiences with threat modeling within the organization? These research questions were determined in collaboration with the sponsor, ensuring that the study addressed relevant and meaningful aspects of the subject matter. However, the subsequent research was conducted independently, safeguarding the objectivity and impartiality of the findings and conclusions. The sponsor was provided with a report of the study findings [29], on which this paper is based.

The goal of this research is to provide qualitative insights into the state of practice of threat modeling. Hence, this paper refrains from using precise numbers or percentages when discussing observations, as they would give a false impression of accuracy due to the limited number of interviewees. Future research may aim for a quantitative characterization of the state of practice through a larger set of interviewees.

2.2 Study design

The interview guide [28] was constructed based on the research questions (see Section 2.1) and lists the different topics to discuss during the interviews in the form of questions. The interviews themselves were performed using the technique of responsive interviewing [20], allowing the interviewers to delve into more detail when appropriate. This means that

Table 1: Overview of the organizations

Sector	Focus	Participants
Energy	OT Systems	1
Finance	Software development	4
Marine	IT Infrastructure	1
Public sector	Software development, advice	3
Transport	Software development	4

the questions from the interview guide were not asked literally nor sequentially; interviews took the form of a natural conversation, merely guided by the topics to discuss.

Ethical approval for this research was obtained from KU Leuven's ethical committee[1] before potential participants were contacted. All interviewed participants have signed an informed consent form. After an interview, each participant was offered a 20 euro gift voucher for their participation.

2.3 Recruitment process

Given the study's focus on organizations that are part of the target audience of the sponsor (NCSC), the sponsor provided a list of contacts at the relevant organizations to reach out to. As the goal of our research was to gain insight into the state of practice, we informed these contacts that potential interviewees should be directly involved in threat modeling. All contacts received from the sponsor agreed to an interview and/or provided other contacts within their organization; we unfortunately have no information on contacts that the sponsor approached and declined, or their reasons. Potential participants with the relevant expertise were provided with an informed consent form [27] and information sheet [26] which described, among others, the goal of the study, the methods used, information on voluntary participation and withdrawal, compensation, potential risks, confidentiality, data processing, and contact information of the researchers.

In total, 13 participants from 7 organizations agreed to participate, resulting in 10 interviews (in three interviews, two participants were interviewed at the same time). Each organization has thousands of employees, and all participants have a role dedicated to security. General characteristics of the participants and organizations are provided in Table 1.

2.4 Data collection process

Two interviews were conducted at the participant's offices; the others were conducted online (through a video call). The interviews were performed by the authors (KY, LS, and SV). One researcher took the lead during the interview. For early interviews, other researchers observed and took notes for consistency with later ones. Each interview lasted approximately

---

[1]KU Leuven Social and Societal Ethics Committee (SMEC), case ID G-2021-4578-R2

one hour, except for joint interviews (approx. 90 minutes). All interviews were in Dutch, except one (in English).

After a brief introduction and repeating the agreements on confidentiality and data protection, the remainder of the interview was recorded (using a microphone for on-site interviews, and the built-in Teams functionality for online interviews). During or after the interview, some participants also briefly showed reports of threat models on which they have worked.

Afterwards, the recordings were transcribed using automated transcription, the results of which were subsequently checked and corrected using the original recording. Automated transcription was initially performed using the built-in functionality of Microsoft Word; for later interviews, a fully offline implementation of Whisper [8] was used.

The interview transcripts were subsequently anonymized manually, by replacing or scrubbing all information that would enable identification of the participant or the organization. All participants received a copy of the anonymized transcript of their interview, and had the opportunity to add remarks, provide corrections, or highlight potentially identifying information. Two participants explicitly confirmed that the information in the transcript was still accurate; one other participant clarified changes in the organization that occurred after the interview. All copies of recordings, non-anonymized transcripts, and notes are destroyed at completion of the research study.

2.5 Analysis procedure

The analysis of the research data involved a systematic coding process of the anonymized transcripts. To facilitate organized coding, a software package for qualitative data analysis (ATLAS.ti) was employed. The coding process used a mix of bottom-up and top-down codes, allowing themes and patterns to emerge from the raw data [5]. Initial (top-down) codes were generated in alignment with the research questions (for example, related to demographics, process/execution, etc.), augmented with the researchers' recollections from reading through and anonymizing the transcripts. Throughout the coding process, these codes were complemented with (bottom-up) codes that capture significant other concepts, ideas, or phrases relevant to the research questions. After coding, codes with similar meanings or concepts were grouped into higher-level codes. The complete codebook can be found online [25].

Based on the coded transcripts, recurring themes and challenges were identified, for which quotations were collected. For interviews that were conducted in Dutch, the quotations used in this paper were translated to English. These quotations served as evidence to support the findings and conclusions drawn from the analysis. A single researcher (SV) was in charge of coding,[2] identifying themes among the coded transcripts was done by multiple researchers (KY, LS, and SV).

---

[2]While we agree that multiple coders would improve the reliability, this is not a crucial aspect of a qualitative surveys according to the ACM SIGSOFT Empirical Standards for Software Engineering [1].

3 Results

This section answers the research questions based on the data collected through the interviews.

3.1 Embedding of threat modeling activities

The first research question concerns the embedding of threat modeling in organizations. This is split up into three sub-questions regarding (1) the definition and perceived benefits, (2) the organizational motivation, and (3) using the results.

RQ1.1. Why do organizations threat model?

All participants agree that threat modeling is an important analysis activity in the development process. Participants frequently mention using threat modeling to analyze and map threats, vulnerabilities, or risks; combined with thinking about potential countermeasures, although some participants note the limited support in this regard. Less frequently mentioned aspects of threat modeling include the importance of considering particular threat actors (*"know your enemy"*), explicitly thinking about key assets (*"what do we want to protect?"*), abuse cases (*"next to the use cases, to also define abuse cases [...] and think what could go wrong in the flow"*), and the supporting role of threat modeling in subsequent activities such as pentesting (*"it is also an excellent basis for [a pentest]"*).

Threat modeling is, however, not always explicitly labeled or systematically executed, and several participants mention security practices being performed in their organization which closely resemble threat modeling without being labeled as such (*"[security practices happen] a lot, but not structurally and not under the umbrella of threat modeling"*).

The main benefits organizations perceive are twofold. First, threat modeling is employed to gain understanding and insight into an application's security concerns (*"to develop more secure products"*). Second, it is also a useful technique to raise the overall security awareness of teams (*"they learn to think about threats"*), and to give the teams a way to talk about security (*"a way for them to discuss information security in a practical way within their team"*). A third goal mentioned by one of the participants is to use threat models as a communication tool for security with non-technical people (*"so that [non-IT] people also get a good understanding of how certain things can occur"*).

RQ1.2. How are stakeholders motivated to threat model?

There is a strong focus on promoting threat modeling internally as a technique for the development teams to apply, rather than mandating threat modeling through organizational policies. Awareness measures range from simply mentioning the technique (*"tell them once, let it simmer"*) to organizing internal workshops (*"so that people at least understand what threat modeling is and why we do it"*).

Most participants stress that development teams should internally recognize the usefulness of threat modeling. This ensures that the motivation comes from within the team (intrinsic motivation) rather than being imposed (extrinsic motivation) (*"The initiative to do [threat modeling] should come from the developers. [...] the moment you start forcing threat modeling, people naturally lose enthusiasm and do it because they have to and not because they see the usefulness and necessity of it."*). In some cases, threat modeling is explicitly required for certain types of applications (e.g., depending on the sensitivity or the business impact). In general, however, it is rarely imposed, as doing so would result in it becoming a checkbox activity (*"once you start having these compliance requirements [. . .] they will just not write stuff down anymore. So, the question is, what is the impact of that going to be?"*).

The relevance and usefulness of threat modeling are already appreciated by teams in several organizations (*"threat modeling is also well received, generally, by the teams"*).

RQ1.3. How are the threat modeling results used?

Follow-up is mostly an ad-hoc activity for which the responsibility usually lies with the team itself (with the exception of some severe issues where the security team actively follows up). How to monitor and follow up on the results more systematically is a recurring challenge (*"That varies depending on the team, and also on the priorities of the product owner [...]"*). This will be explored in more detail in Section 3.4.

One activity that does frequently occur is pentesting, which allows to verify the implementation of mitigations and tends to resurface issues that were not resolved by the teams. Having access to a threat model was mentioned to simplify the pentest process. There is also an opportunity here for positive feedback. Analyses that do not uncover any findings often result in minimal reports, and stakeholders may think that they waste time and resources without really gaining any value. The observation that the team properly implemented the right mitigations is, however, something that can also be actively communicated to them as positive feedback (*"[as a pentester,] it's not really accepted yet that you just go back to a customer, and say, 'gee, you guys just did a great job'."*).

Summary

While there is no consensus on the definition of threat modeling and what these activities specifically entail, all participants recognize and agree on the importance of threat modeling. The obvious benefit perceived by participants is the identification of security threats, as this is the primary reason to perform threat modeling. An important secondary benefit recognized by many participants is raising security awareness among the development teams. Intrinsic motivation of the development teams to perform such analyses was considered an important aspect by many participants, stressing the desire

to have the teams want to perform such activities rather than a mandatory assessment that would be perceived as checkbox compliance exercise. In some organizations, threat modeling is required for critical applications. Using the threat modeling results and especially the more systematic use and follow-up of the results is more of a challenge for organizations.

3.2 Involved organizational roles

The second research question concerns the involved stakeholders, specifically (1) during threat modeling, (2) introducing threat modeling, (3) the goal of management and operations, and (4) the involvement of third parties.

RQ2.1. Who is involved in threat modeling activities?

Promoting threat modeling and making development teams aware of its benefits is mostly done by dedicated security teams. In general, the development teams themselves are responsible to start threat modeling, but the security team may also suggest or mandate threat modeling, especially for high-risk applications (as was described in Section 3.1).

The main stakeholders involved in the threat modeling activities are the development team, the product owners, and an architect, supported by a facilitator from the security team. To a lesser extent, testers, information security officers (ISOs), and operations are involved. The lesser involvement of these other roles is usually the consequence of their limited availability. Two participants mentioned that involving incident response people can be particularly useful, enabling the additional insight into which types of security concerns are relevant and actively abused in incidents; however, their involvement is rare (*"[...] they don't have the capacity [to attend threat modeling sessions]"*, *"we share our threat models with [incident response] [...] but I think it would be better if they just join threat model sessions."*).

RQ2.2. How is threat modeling introduced?

For most organizations, threat modeling has been introduced fairly recently (i.e. in the past 5–6 years) by the security team. Most people that take up an active role in introducing threat modeling to an organization had prior experience with pentesting (*"We noticed that information security officers found it difficult to start up threat modeling activities, and because pentesters are more involved in the [development] activities, we noticed that they could do so more easily"*). In general, participants did not mention a specific trigger to start up threat modeling activities other than having heard about the technique and its benefits. Structured approaches to start up threat modeling activities were also not mentioned: in general, the security team gets familiar with threat modeling through literature (e.g., Shostack [24]) and gradually learns to apply existing threat modeling approaches (*"we gradually learned [to threat model] together"*).

One exception is that, in one of the interviewed organizations, external expertise was consciously attracted to introduce threat modeling into the organization (*"I really followed [hired expert] around for 3 months, almost like a shadow, and that helped a lot too"*), which enabled overcoming organizational challenges and habits (*"[the expert] does not have the bias of the organization and its processes"*).

RQ2.3. What is the role of management and operations?

Information security officers and management positions are often only involved in the communication of the results. However, it is often difficult to communicate these results and clarify the usefulness of threat modeling. Being able to demonstrate a clear business impact and having success stories can help to communicate the results (*"We share successful [threat modeling] stories from time to time, so that [management] sees the added value."*). Management positions are rarely involved during threat modeling sessions. Furthermore, follow-up by management is lacking, and challenging in general (Section 3.4). Operations, including members of the Security Operations Center (SOC), are also rarely involved, except when applications are bought from third parties and need to be integrated. In such a case, operations are the main stakeholder.

RQ2.4. Are any external parties involved?

In all interviewed organizations, threat modeling is performed in-house, with support from the security team. In a single case, however, external expertise was consciously attracted to introduce threat modeling into the organization, which enabled overcoming organizational challenges and habits.

When software is acquired rather than developed in-house, it may be necessary to involve the provider of the application when making a threat model of the integration. Similarly, when software is hosted externally, the host may need to be involved in the threat modeling process. Not all third parties, however, provide equally detailed security documentation (*"Then you depend on, on the one hand, [third parties] being able to provide information, and on the other hand also the level of maturity on security of those kinds of companies."*). Mitigating security threats which require help from the external party is therefore mentioned to be challenging.

Summary

The main stakeholders during a threat modeling session are the development team, the product owner, and an architect, usually supported by a facilitator from the security team to provide expertise. Testers, ISOs, and operations are usually not involved. Management roles are often only involved in the communication of the results. In many cases, the introduction of threat modeling was triggered by prior (pentesting) experience of a security team member. One organization hired

external expertise for this particular purpose, which was well-received. The security team then further propagates threat modeling within the organization.

3.3 Threat modeling process

The third research question concerns the threat modeling process, including (1) the trigger, (2) teaching threat modeling, (3) inputs and models, (4) threat elicitation, and (5) output and follow-up.

RQ3.1. When are threat modeling activities triggered?

As described in Section 3.1, while threat modeling may be mandated for high-risk applications, organizations foster intrinsic motivation, and threat modeling activities are therefore also mostly triggered by development teams that want to investigate the security of their application. This usually involves reaching out to the security team for training or support, or for confirmation or feedback on their threat models.

There is an overwhelming consensus that threat modeling is a continuous effort and thus requires periodic re-assessments. Implementations vary from development teams reaching out for feedback on their models to the security team frequently checking in with developers to do a re-assessment if necessary. In practice, such reassessments, and follow-up in general, depends on the willingness of development teams and the priorities of the product owner, and overall is challenging. When prompted about tool support, participants recognize the opportunity for tooling and automation such as integration in CI/CD pipelines to trigger reassessments if changes may introduce new threats, but none of the organizations do so at the moment, mostly due to the lack of tool support. In general, threat modeling therefore remains mostly a one-time activity, and models are infrequently revisited or updated.

The usefulness of early threat modeling is recognized, but this is in practice not always done. One of the reasons for this is the backlog of high-risk applications which require a threat model, leaving less room for the security teams to support early-stage threat modeling sessions (*"[...] we're actually catching up now first, which means you're mostly threat modeling on applications that are already live"*). Even so, there are several instances were threat modeling was applied very early in the development lifecycle, in tender processes and procurement, leading to valuable feedback and concrete security requirements. For example, in one specific case mentioned by participants, threat modeling during procurement later prevented a specific ransomware attack.

RQ3.2. How is threat modeling taught?

In general, the interviewees indicated that a threat modeling session usually starts with an introduction to threat modeling, which varies from a couple of slides (*"a few slides, two or*

three, to shortly explain the methodology ") to more lengthy ones (*"we first gave an introduction of about 40-45 minutes"*).

Providing separate learning materials or organizing workshops before the actual threat modeling session is also prevalent. While generally perceived as useful, participants indicate that separate learning materials do not suffice to teach teams to threat model independently (*"I don't see teams picking it up and doing this completely independently any time soon"*), and teams may not always go through them (*"I don't think they go through the materials we are sharing with them"*).

Besides introducing the methodology and basics of threat modeling, the following aspects are usually covered during training. First, teaching teams to think about what can go wrong was mentioned several times (*"worst-case thinking really needs to be taught"*). Second, while teams may be more comfortable with a well-defined method, several participants note that the exact methodology in general is of little importance, and that thinking about security at all is more important than following strict guidelines (*"the most important thing is to start [threat modeling]. You can't really do something wrong"*). Third, teams may lack security expertise, so some examples or prevalent threats may also be illustrated. This lack of security expertise was also mentioned as the main reason why teams are not confident to independently start threat modeling (i.e., without the presence of a facilitator or security expert), as for example described by *"My impression is also that they are perfectly capable of doing it themselves if they have seen it once. That last 5% is indeed 'what do we [as security experts] see?'. And they can't do that themselves."*

RQ3.3. What kind of inputs (models) are used?

The first step of a threat modeling session is usually to create a model of the application or system being analyzed. Overall, the diagrams created or used in the context of threat modeling can take various forms, ranging from re-used architectural documentation to whiteboard diagrams. There is a balance between diagram quality conventions and the effort for teams to adhere to them because of the overhead they introduce. As a result, tool support for creating diagrams is mostly limited to drawing tools like Threat Dragon [18], but in some cases more elaborate modeling support like Microsoft's Threat Modeling Tool [13] is also used. In terms of model types, data flow diagrams (DFD) were most commonly used for software systems. One exception is the interviewed organization focused on operational technology (OT), which used a map of the network layout as the primary model.

A broadly recognized benefit of threat modeling is that it forces the explicit consideration of architectural documentation which can be either non-existent or, more frequently, outdated. Threat modeling therefore provides an incentive to revise and update this documentation. In some cases, the security teams construct initial diagrams to bootstrap the threat modeling activities, based on the inputs of the development teams. An important concern for the creation of the diagrams is the scope of the analysis to ensure a focused discussion.

RQ3.4. How are threats elicited?

STRIDE (a mnemonic for *spoofing, tampering, repudiation, information disclosure, denial of service, and elevation of privilege*, which can be used to guide threat modeling exercises) is most frequently mentioned as the main driver for threat elicitation. Threat elicitation is not necessarily performed systematically (e.g., using the STRIDE threat mapping table as described in Shostack [24]). Indeed, organizations prefer flexibility, giving development teams freedom in how to do the analysis. Other approaches such as PASTA [35] are used when the need arises (*"we chose STRIDE at the time mainly because it's very easy to explain and very accessible"*).

Besides the system model, inputs that are frequently leveraged during threat analysis are the ingress points in the system, attack vectors, types of adversaries, and attack scenarios (*"[...] which threats, and which attackers do we think are interesting?"*). Organizations want to reuse any such organization-specific knowledge across multiple analysis activities.

Finally, participants perceive the value to be mainly in the process rather than in the quality of a threat model. That is, it is more important to do the analysis than to have a detailed threat model (*"Going through the process is perhaps the most fruitful."*). There are also generally no strict criteria on when analyses are finished. Usually, sessions end naturally when no new threats arise or when all elements have been covered.

RQ3.5. How are results reported and tracked over time?

In general, threat modeling results in a report containing the system model, identified threats, present mitigations and recommendations to resolve unmitigated threats. In some cases, richer descriptions are made using attack scenarios. To reduce the number of issues to tackle, threats can be prioritized (*"[...] a summary of the relevant risks, at the basis of which recommendations are made"*).

Overall, organizations want to limit the reporting overhead as writing everything out in textual reports requires substantial effort with limited returns (*"writing takes a lot of time, and I don't know if it's always worth the effort."*). In some cases, presentations of the results are used to limit such overhead. The execution of the threat modeling process itself is considered more important than the reporting. While tool support is considered, linking the findings to business risks remains a challenge and requires manual effort.

Follow-up is mostly an ad-hoc activity for which the responsibility usually lies with the team itself (with the exception of some severe issues where the security team actively follows up). How to monitor and follow up on the results more systematically is a recurring challenge (*"That varies depending on the team, and also on the priorities of the product owner*

[...]"). This will be explored in more detail in Section 3.4. While going through the process to create security awareness is, in some cases, the main goal, some participants expressed a wish for more frequent and standardized follow-up, but strict policies may not be favorable and result in compliance-like checkbox activities.

Summary

Threat modeling sessions are mostly triggered by development teams wanting to examine the security of their system or application. Participants agree that threat models should be started early on in the development lifecycle and require periodical reassessments, but this is not common practice as security teams are currently prioritizing a backlog of high-risk, operational systems. While dedicated training sessions are both commonplace and essential for instilling the proper mindset, enabling a team to independently execute threat modeling can be challenging. Software models used during threat modeling take various forms ranging from free-form whiteboard drawings to structured notations like data flow diagrams. (Up to date) architectural models are not always available for re-use, so (re)constructing them becomes an important part of threat modeling. Concerning the use of models, pragmatism prevails over conforming to standardized notations. A pragmatic use of the STRIDE acronym is the most common approach for identifying threats during threat modeling. In this context, taking action and moving forward is considered more valuable than achieving a perfect threat model or prioritization of threats. In most organizations, no strict follow-up processes for the results of threat modeling are in place.

3.4 Threat modeling experiences

The fourth and final research question concerns experiences with threat modeling, including (1) positive experiences, (2) challenges and (3) causes of difficulties.

RQ4.1. What are positive threat modeling experiences?

A major success experience consists of teams becoming increasingly aware of security and the advantages of threat modeling. In some cases, these insights directly prevented concrete attacks (i.e., ransomware attacks). Furthermore, threat modeling is mentioned to decrease the effort required to develop pentests. Participants also indicate that teams are starting to threat model earlier in the development lifecycle, and do so more periodically, which has a positive impact on the complexity and duration of threat modeling sessions. Threat modeling during the design phase, although not prevalent, was also indicated to be beneficiary, leading to concrete security requirements which can be taken into account throughout the remainder of the development lifecycle. Finally, involving external parties to introduce teams and organizations to threat modeling was also indicated to be beneficial.

RQ4.2. What are threat modeling challenges?

Threat modeling challenges described by participants relate to (1) planning, (2) training materials, (3) modeling, (4) threat elicitation and prioritization, (5) follow-up, (6) tool support, (7) involving management, (8) demonstrating effectiveness, and (9) intra-organizational differences. A comprehensive overview is provided in what follows.

Planning. Scoping threat modeling activities is crucial to manage their size and complexity. Starting too early may lead to an ill-defined scope, starting too late to a too large scope. Several participants described difficulties finding the right time to start or revisit a threat model. Furthermore, mitigating issues, especially design issues, may be difficult or even impossible when applications are already fully implemented or deployed (*"what can you still do, right?"*).

Security teams themselves may also experience difficulties to plan a session if teams request it close to their deadline (*"Not all teams are aware of our schedule as [the security team][...]"*). A more general challenge is that security teams simply may not have the resources to provide threat modeling support to all teams (*"we simply don't have the capacity for that yet, because we just have so many development teams."*).

Regarding the duration of threat modeling activities, teams may lose interest if a session takes too long, especially if it is dominated by one or a few people, or gets too technical, and teams may be reluctant to start threat modeling a large system or application due to the amount of time that must be invested (*"You have to keep the focus time short, right? [...] Otherwise the team gets bored or there's no time left."*).

In general, participants described that the best way to tackle planning-related challenges would be to make threat modeling a part of the default workflow of the teams, as they themselves know best when threat modeling would be opportune.

Training materials. One participant mentions that creating worked examples for threat modeling is challenging, both because it is time-consuming (*"they tend to be very time intensive to actually create"*), and because teams tend to focus on the specific material covered in the examples, which may hinder them from finding other issues (*"[...] the only thing they're going to be doing is regurgitating the exact same thing that you told them during the training, at which point, yeah, you can also just give them a checklist"*). Another participant mentions the lack of real-world experiences on how to introduce threat modeling to an organization (*"you rarely hear about, well, I did it this way, and you need this, and you need these contacts, and you need to arrange it this way."*).

Modeling. Architectural documentation is seldomly available or up-to-date (*"the documentation we get is almost never up-to-date"*), which hinders the creation of models and diagrams (*"the fact that we have to spend the beginning of a*

session on getting the model correct, or as correct as possible is, in my view, a bit of a waste of time"). An underlying problem is that a single comprehensive overview is usually not available (*"there is no single record, with* the *truth, not even on a conceptual level"*). Tackling this issue by involving multiple architects was also mentioned not to be favorable by one participant, as this may lead to lengthy discussions (*"we prefer to have only the architect who is most involved there."*).

Regarding the types of diagrams used, one participant describes that data flow diagrams may not be ideal for more specific and technical types of analyses (*"For more the protocol related things, for example, this is where it kind of, kind of breaks down [...] because you really want to look at much more specific and technical issues."*).

Threat elicitation and prioritization. Participants prefer to choose a methodology and stick to it to avoid losing time on discussions (*"If you aren't careful, you will have a lot of discussions about the form before you actually get started."*). Specifically for STRIDE, one participant mentioned that it does not scale well, as even for smaller applications, the amount of threats may rise rapidly (*"as the number of flows in and out of an application increases, the amount of time you have to spend on it increases exponentially"*). As a result, applying STRIDE during more agile workflows was indicated to be cumbersome (*"in an agile sprint or something like that, STRIDE is quite a cumbersome method"*).

Regarding risk estimation, it requires both security expertise and domain knowledge, and guidelines on how to do so are lacking in general (*"First, we don't provide a clear framework, how to do that themselves, and second, even if we had some way to evaluate the risk, they would still be guessing it, it's not going to be accurate enough."*).

Other related challenges include not thinking about the attacker (*"knowing who you're up against... I notice that a lot of people don't talk about that"*), approaching a threat model too much from a pentest point of view, which may lead teams to get stuck on the details (*"[sometimes] we treat the threat model a little too much as a starting document for our pentest, rather than a standalone thing"*), communication (*"Totally different sides of an organizations are suddenly going to be collaborating [...] Purely on language alone, you have to be very careful with that."*), and supply chain management (*"[...] yes, we are fine, but what about our suppliers?"*).

Follow-up. In general, systematic follow-up on the outcome of threat modeling sessions is lacking. Security may not be a priority of the team or product owner, which may lead to threat modeling outputs being ignored (*"our product owner doesn't think that's exciting enough right now"*). This is especially the case when threat modeling is mandated by some policy (*"they just want a list, and ticked off, and then you've done well"*). Participants do agree that this is not due to the lack of security interest, but rather because teams have limited time

(*"It's not that they don't want to do security, but they have so many other things to think about besides security."*).

Following up was mentioned to be difficult for multiple reasons. First, acting on the threat modeling results may require the help of external people, for example for externally hosted applications. In such cases, it may take time to get this on the agenda of the external entity (*"To solve an issue [with an external host] would involve creating a ticket, and most likely lengthy email conversations, phone calls, ... "*). Furthermore, as mentioned in Section 3.4, threat modeling sessions are planned late in the development cycle in some cases, which limits the changes that can be made to an application (*"[...] and then we find out that there are actually quite insurmountable problems in the software"*).

Participants also described that follow-up is challenging if it involves other teams or stakeholders within the organization. For example, there is a risk of interfering with previously made (design) decisions, potentially taken by other teams (*"[...] they all take separate, siloed actions and don't take into account what preceded it, or too late."*). This is especially relevant when there is a business incentive to deploy as soon as possible. In such cases, deciding what to do or how to process the output of a threat modeling session (if at all) may become tedious and time-consuming (*"that generates a lot of discussion"*). Furthermore, even if teams want to take into account the threat modeling outcomes, interpreting the results was indicated to be challenging by the majority of the participants (*"It might be a problem with other teams interpreting threat models, one team interpreting a threat model [differently from] another team."*). Standardizing the outputs may be one way to tackle this challenge, but too much standardization may deter teams from threat modeling at all (*"[...] then you do get some interchangeability of [threat modeling results], without immediately killing the whole enthusiasm by putting it in a straitjacket, because that's not the goal either."*). Another challenge related to system models is a lack of diagram conventions, which inhibits the use and interpretation of threat modeling documents by other teams. Finally, one participant describes the risk of assuming that other stakeholders will take care of an issue (*"Assuming that another team does something [... is] more a problem than having the same circles, squares, arrows and whatnot."*).

Tool support. Tool support (e.g., Microsoft's Threat Modeling Tool [13] and Threat Dragon [18]) was indicated not to be user friendly (*"I find that it lacks some things in terms of usability"*). Microsoft's Threat Modeling Tool specifically was mentioned to require a lot of detailed inputs in order to get to useful output (*"you really have to fill out a lot to get useful information"*; *"you also don't want to tire the team with all those details, like, what TLS version are you using, and stuff like that"*). Interpreting the output of threat modeling tools was also indicated to be challenging, mainly because it requires security expertise (*"at the very least you want to pre-*

vent [the teams] from, yes, not having the knowledge and, yes, then simply disregarding [the output]"). For these reasons, except to draw simple diagrams, using threat modeling tools during a session was generally avoided.

One participant mentioned that, to make threat modeling tools a part of the general workflow of teams, they should be simplified (*"a simple implementation so that teams can start using it at all"*). Another issue mentioned by one of the participants is that threat modeling tools do not allow to model business logic well (*"it's not really very easy yet to include business logic"*). Finally, while participants indicated that integrating threat modeling tools in a CI/CD pipeline could be beneficial, none of them do so at the moment (*"I don't see how you could integrate threat modeling specifically into your CI/CD pipeline."*). One participant described the idea to automatically create tickets for threats, but due to the number of threats that are identified by threat modeling tools, this could also be challenging (*"[to] have ten thousand tickets automatically open... That's not going to be nice."*).

Involving management. (Risk) management may not always be aware of the added value of threat modeling, which makes getting support, time, and resources for threat modeling challenging (*"Getting resources to do it from the higher-ups, that always requires work."*). Ideally, according to one participant, management should not push or mandate threat modeling, but support teams wanting to do it (*"I would hate to have to push that from a leadership role. [...] But management, according to me, does play a role in accepting it, seeing the added value of it and being able to translate that back to their stakeholders."*).

Second, involving management during threat modeling sessions could provide useful insights, but is challenging for two reasons. First, management may not be aware of the benefits of them being present and may think that threat modeling sessions require a strong technical and/or security background (*"They are very quickly afraid that it really becomes a very technical session."*). Second, management simply may not have the time to join threat modeling sessions (*"[...] we have a single ISO right now. [...] Yeah, that's too few."*).

Finally, management does not follow up on the results of threats modeling sessions according to several participants (*"that just doesn't always happen or, at least, not consistently"*). Even if management would like to follow up, they may not always be able to correctly interpret threat modeling reports, because they are not always involved or familiar with the context (*"You need to be able to interpret a report."*). This lack of follow-up could result in a lack of oversight across applications and an organization in general (*"that leads to lack of oversight, where you can miss things"*).

Demonstrating effectiveness. Measuring the effectiveness of threat modeling, and security in general, is indicated to be challenging (*"evaluating whether threat modeling helps*

to achieve security is very hard, because you can't really measure security", *"it's an article of faith and we are part of the threat modeling church"*). However, in order to create awareness and motivate teams to do threat modeling, being able to communicate its added value may be crucial (*"What is the added value of threat modeling, right? And I think, making that clear and communicating unambiguously [and] empirically backed up [...] will be decisive."*). One participant mentions that the results of a pentest could be a starting point to evaluate a threat model, for example to identify issues that were missed during the threat modeling session (*"[...] does the pentest show up stuff that wasn't in a threat model or assumption that were incorrect?"*). Evaluating the artifacts created and used during a threat modeling session itself is also indicated to be challenging (*"Looking at the artifacts themselves [...] that's also an area that's still a bit open."*).

Intra-organizational differences. While our interviews only include one participant with a focus on OT (including for example industrial control systems), an important source of difficulties for that participant stems from the inherent (cultural) differences between the IT and OT domain. Mitigating certain threats or creating more secure systems may involve enforcing policies (for example related to patching), also on the OT side, even though IT policies don't always translate well to an OT context (*"IT organization as I know them are often quite bold and understand little of the OT, yet they feel we must comply with their policies."*). Understanding the differences between IT and OT, and effective communication between both sides, is therefore seen as an important but challenging aspect of security in general (*"embrace the fact that our worlds are different"*).

RQ4.3. What are the causes of the experienced challenges?

Challenges concerning motivation, timing, and follow-up are mainly caused by product owners, information (security) officers, and other management roles not being aware of the benefits of threat modeling. A root cause for this is that demonstrating the effectiveness of threat modeling is challenging. Teaching teams how to do threat modeling is furthermore complicated by a lack of a security mindset and knowledge within the team. Finally, the limited use of software tools for threat modeling is due to the required effort that outweighs the perceived benefits.

Summary

Development teams in the interviewed organizations are becoming increasingly aware of threat modeling and its advantages, and teams are starting to threat model earlier in the development lifecycle, and more periodically, which has a positive effect on the complexity and duration of threat modeling sessions. Other positive threat modeling experiences

mentioned by participants include the prevention of concrete attacks, the use of threat modeling results when pentesting, and involving external parties to help introduce threat modeling to an organization.

Threat modeling related challenges faced by organizations include (1) finding the right time to start a threat model and finding a time slot that fits all stakeholders, (2) dealing with the overall lack of security expertise when introducing teams to threat modeling, (3) overhead during threat modeling sessions related to, among others, the lack of architectural documentation, discussing and deciding on the methodology, risk estimation, and long technical discussions, (4) the lack of follow-up, adequate tool support, and management involvement, (5) demonstrating the effectiveness of threat modeling, and (6) different (security) cultures between different parts of the organization, and IT and OT in particular. A lack of (1) threat modeling awareness at the level of product owners and management roles, (2) security knowledge among development teams, and (3) adequate tool support have been mentioned by organizations as potential causes of these challenges.

4 Discussion

This section discusses the main implications of our observations for practitioners, potential directions for future research, and the limitations and threats to validity of our study.

4.1 Advice for practitioners

Based on this study's findings, the main advice for organizations is to consider and incentivize thinking about security in any shape or form, rather than mandating threat modeling and imposing strict requirements on the methodology. Indeed, one of the major perceived benefits by participants is that it increases security awareness among the development teams. When evaluating their threat modeling practices, it is therefore important for organizations to recognize that there is no one-size-fits-all threat modeling approach that has worked for every organization, and that even within a single organization, different teams or applications may require a different approach. In this regard, forcing the use of a specific tool with the hopes of it leading to an efficient and fruitful threat modeling process should be avoided. Indeed, most of the interviewed organizations tried to use or considered using tool support to (partially) automate threat analysis, to support the creation of software models, or for more systematic follow-up, but adequate tool support seems to be lacking. Especially for organizations that are yet to start or just introduced activities related to threat modeling, it seems that successful instantiations of threat modeling spring from giving some space and flexibility to the development and security teams to see if, where, and how threat modeling can provide value, and gradually building upon and expanding this expertise.

In an ideal scenario, threat modeling is done early in the development lifecycle, as mitigating discovered threats in large, existing systems that are already operational may not be straightforward. Furthermore, threat modeling should ideally be repeated when changes are made (e.g., new features, or changes to the architecture). However, several organizations have highlighted difficulties with planning and finding the right time to threat model. Making threat modeling a part of the default workflow of development teams may alleviate such challenges, yet care must be taken that it does not become a checkbox activity. The fact that threat modeling allows gaining and maintaining a mutual understanding of an application and its architecture can also be promoted to incentivize teams to periodically apply threat modeling.

Product owners and management roles in general need to be aware of the potential benefits of threat modeling and allow for the necessary time for development teams to learn and apply this skill. Therefore, besides incentivizing development teams, awareness campaigns aimed at management roles could be fruitful. Such raised awareness may also contribute to better follow-up of threat modeling results which, in many of the organizations, appears to be limited and ad-hoc. Besides following up on threat models, actually involving management roles during threat modeling sessions was indicated by participants to be valuable, yet care must be taken that such sessions then do not become too technical.

Finally, organizations could use the research questions of this study as a starting point to evaluate their own threat modeling practices.

4.2 Directions for future work

The findings of this study reveal potential directions for future research regarding threat modeling. First, in order to further convince management roles of the benefits of threat modeling, the effectiveness and return on investment of threat modeling could be investigated, be it in terms of finding threats, raising security awareness (and thus preventing future threats), or supporting subsequent security activities like pentests.

Second, participants recognize the potential benefit of using tool support to automatically trigger re-assessments of threat models when significant changes are made to a system, but currently available threat modeling tools do not offer such capabilities. Future research and development efforts could aim to improve tool support and allow such integration in a CI/CD pipeline. Furthermore, the usability of threat modeling tools should be investigated, as participants agree that currently available tools require too much effort and, as a result, are not fit to be integrated in agile development processes.

It should, however, be noted that the described usability issues with current threat modeling tools may not necessarily be encountered by organizations that have heavily automated their threat modeling activities, and which may appreciate more detailed modeling capabilities and outputs. Still, since

the interviewed organizations utilized little tool support, it would be interesting to see how threat modeling tools could be refined to support such organizations, for example by guiding development teams through a threat modeling session without the presence of a facilitator of the security team, which was mentioned to be difficult mostly due to the lack of security knowledge among developers.

Finally, rather than such 'user friendly' tools (and frameworks in general) not being available, another issue could be that such tools exist, but practitioners simply do not know about them, or do not know how to use them. A similar phenomenon was investigated by Canedo et al. [4], who describe that privacy requirements elicitation tools and techniques used and studied in literature do not align with the ones used in practice, partially due to the lack of dissemination and training materials. Participants in our study also described that their choice of using STRIDE over other methodologies is partly due to there being more training materials available for STRIDE. Therefore, future work could investigate practitioner needs in terms of training materials, and how novel threat modeling techniques could be better disseminated.

4.3 Threats to validity

This study is based on only a few organizations (13 participants in 7 organizations), where often only one person from each organization was interviewed. Although this person was always well-placed and had a comprehensive view on threat modeling in the organization (i.e., a member of the organization-wide security team), they may not be fully aware of all threat modeling initiatives.

This study is also subject to several selection biases. First, it is performed on target organizations of the NCSC, which typically are large organizations in critical sectors with a dedicated security team; software development is not their main activity. The results are thus not necessarily representative for other (smaller or commercial) organizations. Furthermore, regarding self-selection bias, the organizations already implement some form of threat modeling and are willing to openly talk about it, and contacts were provided by the sponsor of this research. Moreover, most of the interviewees are threat modeling 'advocates', appreciating its value, and actively pushing its use. This study does not include (nor encountered) any organizations that have tried and abandoned threat modeling, or where no threat modeling program is being developed.

Interviews being the only research method used, there is a possibility for respondent or social desirability bias (e.g., idealized, or exaggerated versions). Some interviewees showed threat modeling reports of projects in which they participated to illustrate what was said, which partially tackles this bias regarding the findings related to process and outcomes. Furthermore, based on the numerous challenges and negative experiences listed by participants, it is unlikely that an idealized version was presented. Moreover, with a limited interview

duration of one hour (or 90 minutes if two participants were interviewed simultaneously) and the use of a responsive interviewing style, not all topics listed in the interview guide were explored in equal depth in each interview. Potential interviewer bias was reduced by formulating neutral, open-ended questions in the interview guide.

Finally, this study focuses on activities under the name of 'threat modeling'. Other organizations may perform similar activities under a different name (e.g., a security design review, security risk assessment, or the creation of abuser stories). A broader study that focuses on all design-level security activities would be needed for a more complete picture.

The main observed success factors (e.g., fostering intrinsic motivation and pragmatism) and challenges (lack of architectural documentation, follow-up, etc.) are shared by all interviewed organizations. Later interviews revealed no major new or contradictory observations. While this is not a grounded theory study, this indicates a certain level of data saturation. It should be noted that all contacts received from the sponsor (or other contacts provided by them) were interviewed, and that data collected stopped due to the contacts being exhausted, not because data saturation was reached. Still, we are confident that the observations described in this paper will, in general, also apply to other large organizations which are not primarily software development organizations, but have in-house software development teams, and apply some form of threat modeling. Further research is needed on the applications of threat modeling in other types of organizations, notably those focused on OT, as our study only included one participant of that sector.

5 Related Work

This section summarizes studies similar to this one, and highlights findings which differ from our observations.

Several practitioners have described their experiences and lessons learned from applying threat modeling within their organization. For example, Shostack [23] describes the threat modeling approach used by Microsoft, Ingalsbe et al. [9] describe their experiences at Ford, and Dhillon [7] elaborates on threat modeling at EMC Corporation (now Dell EMC).

Additionally, several empirical studies investigated specific threat modeling techniques. First, Stevens et al. [31] introduced a specific threat modeling framework to New York City Cyber Command and report the adoption and efficacy of threat modeling practices. Their participants stated they perform threat modeling in their daily efforts [31], observing analogous awareness benefits as observed in our interviews. Second, Soares Cruzes et al. [30] performed a case study on the adoption of STRIDE in a company comprising five agile development projects and identify challenges similar to the ones observed during our interviews. Third, Bernsmed et al. [2] bundle the results from four different studies on threat modeling as applied in agile projects, focused specifically on

the use of data flow diagrams, STRIDE, and Microsoft's Threat Modeling Tool [15]. Related to the overall organization of threat modeling activities, their observations also include that developers are the main stakeholders, and that there is a need for better integration of threat modeling activities in the development pipeline. Fourth, Weir et al. [36] propose a design for so-called security interventions, which are similar to threat modeling sessions, and evaluate their effectiveness in terms of increased security engagement from product managers and the ability for developers to produce threat assessments. Finally, Trentinaglia et al. [33] describe experiences and lessons learned through conducting threat modeling workshops with practitioners in multiple domains.

As already mentioned, the above-mentioned studies [2, 30, 31, 33, 36] consider specific threat modeling approaches, and mostly focus on the application of the approach. In contrast, our study is not limited to specific approaches, and considers, besides the execution of threat modeling activities, organizational and human-centered aspects including motivation, planning, and stakeholder involvement.

Jamil et al. [11] consider a similar, broad perspective on the organization of threat modeling activities in practice, specifically for cyber-physical systems, through interviews with security experts from several different domains. Contrary to our observations, which mainly relate to IT rather than OT, they describe that the security team executes threat modeling activities separately, using input from other stakeholders (e.g., developers and architects), and that the developers themselves are not actively involved during the process. This may be attributed to IT people not being familiar with the physical aspects of cyber-physical systems, which Jamil et al. [11] describe to be difficult, similar to our findings (Section 3.4). However, if developers are not actively involved during threat modeling activities, the benefit of increased security awareness among developers, which was observed to be one of the main goals of threat modeling by our study as well as related work (e.g., [31, 33]), will not be attained.

A final category of related work is papers which evaluate the effectiveness of threat modeling techniques through experiments in more controlled settings. For example, Scandariato et al. [21] summarize the results of several empirical studies related to threat modeling, including evaluations of STRIDE with respect to the amount and validity of threats found, and a comparison between visual and textual approaches, Tuma et al. [34] evaluate two variants of STRIDE in terms of the number of high-priority threats identified, and de Gramatica et al. [6] investigate if the use of catalogs of threats and mitigations has an effect on the actual and perceived usefulness of security risk assessment methods. While participants in our study indicate that such evaluations could prove useful, for example to convince management of the benefits of threat modeling (Section 3.4), they do not investigate the actual adoption and organization of threat modeling in practice.

6   Conclusion

This paper described the results of a qualitative interview study into the threat modeling state of practice within 7 large Dutch organizations. In terms of organizing threat modeling activities, organizations tend to foster an intrinsic interest in threat modeling rather than putting strict policies in place. The goals for threat modeling are to find and mitigate security threats, but also to raise the overall security awareness among developers. Following up on threat modeling results is indicated to be challenging.

The main stakeholders of threat modeling activities are the development team, an architect, and a facilitator from the security team. Testers and operations are usually not involved, even though their input may be valuable. When software is acquired and integrated rather than developed in-house, however, operations are usually the main stakeholder, and input from vendor may be needed to ensure a secure integration.

In general, a threat modeling session starts with a facilitator from the security team who provides an introduction of threat modeling, including an overview of the methodology (usually based on STRIDE). Then, a model of the system is constructed, the form of which ranges from whiteboard drawings to structured notations like data flow diagrams. Constructing a model may be time-consuming if architectural documentation is lacking. This model is subsequently analyzed, typically in a pragmatic manner. After the session, the facilitator creates a report which is distributed to the stakeholders. Follow-up is mostly ad-hoc, except when critical issues are identified. In general, this is a one-time activity, although participants agree that there should be periodic reassessments.

Positive experiences include the prevention of concrete attacks (albeit seldomly), and (much more commonly) increased developer security awareness. Challenges relate to, among others, planning, training, model creation, risk estimation, and follow-up. These are (at least partially) associated with product owners and management roles not being aware of the benefits of threat modeling, as well as the security team lacking the capacity to assist all the development teams.

Organizations can use these results to help inform decisions to start or extend their threat modeling efforts. Furthermore, threat modeling facilitators and researchers may base future efforts on the challenges identified in this study.

Acknowledgments

References

[1] ACM SIGSOFT. Empirical Standards for Software Engineering: Qualitative Surveys (Interview Studies). https://www2.sigsoft.org/EmpiricalStandards/docs/standards?standard=QualitativeSurveys.

[2] Karin Bernsmed, Daniela Soares Cruzes, Martin Gilje Jaatun, and Monica Iovan. Adopting threat modelling in agile software development projects. *Journal of Systems and Software*, 183:111090, January 2022.

[3] Zoe Braiterman, Adam Shostack, Jonathan Marcil, Stephen de Vries, Irene Michlin, Kim Wuyts, Robert Hurlbut, Brook S.E. Schoenfield, Fraser Scott, Matthew Coles, Chris Romeo, Alyssa Miller, Izar Tarandach, Avi Douglen, and Marc French. Threat modeling manifesto. https://www.threatmodelingmanifesto.org/, 2020.

[4] Edna Dias Canedo, Ian Nery Bandeira, Angelica Toffano Seidel Calazans, Pedro Henrique Teixeira Costa, Emille Catarine Rodrigues Cançado, and Rodrigo Bonifácio. Privacy requirements elicitation: a systematic literature review and perception analysis of it practitioners. *Requirements Engineering*, 28(2):177–194, Jun 2023.

[5] Victoria Clarke and Virginia Braun. Thematic analysis. *The Journal of Positive Psychology*, 12(3):297–298, 2017.

[6] Martina de Gramatica, Katsiaryna Labunets, Fabio Massacci, Federica Paci, and Alessandra Tedeschi. The role of catalogues of threats and security controls in security risk assessment: An empirical study with atm professionals. In Samuel A. Fricker and Kurt Schneider, editors, *Requirements Engineering: Foundation for Software Quality*, pages 98–114, Cham, 2015. Springer International Publishing.

[7] Danny Dhillon. Developer-Driven Threat Modeling - Lessons Learned in the Trenches. *IEEE Security & Privacy*, 9(4):41–47, 2011.

[8] Georgi Gerganov. whisper.cpp. https://github.com/ggerganov/whisper.cpp, 2023.

[9] Jeffrey A. Ingalsbe, Louis Kunimatsu, Tim Baeten, and Nancy R. Mead. Threat modeling: Diving into the deep end. *IEEE Software*, 25(1):28–34, 2008.

[10] IriusRisk. IriusRisk, 2021. https://www.iriusrisk.com/.

[11] Ameerah-Muhsinah Jamil, Lotfi Ben Othmane, and Altaz Valani. Threat modeling of cyber-physical systems in practice. In Bo Luo, Mohamed Mosbah, Frédéric Cuppens, Lotfi Ben Othmane, Nora Cuppens, and Slim Kallel, editors, *Risks and Security of Internet and Systems*, pages 3–19, Cham, 2022. Springer International Publishing.

[12] Peter Mell, Karen Scarfone, and Sasha Romanosky. Common Vulnerability Scoring System. *IEEE Security Privacy*, 4(6):85–89, November 2006. Conference Name: IEEE Security Privacy.

[13] Microsoft. Threat Modeling Tool, 2023. https://aka.ms/tmt/.

[14] Microsoft. What are the microsoft sdl practices? https://www.microsoft.com/en-us/securityengineering/sdl/practices, 2023.

[15] Microsoft Corporation. Microsoft threat modeling tool.

[16] NIST. Secure Software Development Framework (SSDF) Version 1.1: Recommendations for Mitigating the Risk of Software Vulnerabilities (SP 800-218. https://csrc.nist.gov/Projects/ssdf, February 2022.

[17] OWASP. OWASP Top 10 - 2021. https://owasp.org/Top10/, 2021.

[18] OWASP. Threat Dragon, 2021. https://owasp.org/www-project-threat-dragon/.

[19] OWASP. Software assurance maturity model. https://owaspsamm.org/, 2022. Version 2.0.3.

[20] Herbert J. Rubin and Irene S. Rubin. *Qualitative Interviewing: The Art of Hearing Data*. Sage, 2011.

[21] Riccardo Scandariato, Federica Paci, Le Minh Sang Tran, Katsiaryna Labunets, Koen Yskout, Fabio Massacci, and Wouter Joosen. *Empirical Assessment of Security Requirements and Architecture: Lessons Learned*, pages 35–64. Springer International Publishing, Cham, 2014.

[22] Bruce Schneier. Attack trees. *Dr. Dobb's journal*, 24(12):21–29, 1999.

[23] Adam Shostack. Experiences threat modeling at microsoft. In Jon Whittle, Jan Jürjens, Bashar Nuseibeh, and Glen Dobson, editors, *Proceedings of the Workshop on Modeling Security (MODSEC08), Toulouse, France, September 28*, volume 413 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2008. https://ceur-ws.org/Vol-413/paper12.pdf.

[24] Adam Shostack. *Threat modeling: Designing for security*. John Wiley & Sons, 2014.

[25] Laurens Sion, Stef Verreydt, and Koen Yskout. Codebook. https://figshare.com/s/7dcdefa2cf15ee2e01a0, 2023.

[26] Laurens Sion, Stef Verreydt, and Koen Yskout. Information sheet. https://figshare.com/s/b9d3e0f6a821591bba1e, 2023.

[27] Laurens Sion, Stef Verreydt, and Koen Yskout. Informed consent firorm. https://figshare.com/s/3036fb6087838e9770b8, 2023.

[28] Laurens Sion, Stef Verreydt, and Koen Yskout. Interview guide. https://figshare.com/s/4768b946e59ea933cff1, 2023.

[29] Laurens Sion, Stef Verreydt, and Koen Yskout. Threat modeling in nederlandse organisaties. https://www.ncsc.nl/documenten/publicaties/2024/mei/7/index, 2023.

[30] Daniela Soares Cruzes, Martin Gilje Jaatun, Karin Bernsmed, and Inger Anne Tøndel. Challenges and experiences with applying microsoft threat modeling in agile development projects. In *2018 25th Australasian Software Engineering Conference (ASWEC)*, pages 111–120, 2018.

[31] Rock Stevens, Daniel Votipka, Elissa M. Redmiles, Colin Ahern, Patrick Sweeney, and Michelle L. Mazurek. The battle for new york: A case study of applied digital threat modeling at the enterprise level. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 621–637, Baltimore, MD, August 2018. USENIX Association.

[32] Tarandach, Izar. Pytm, 2020. https://github.com/izar/pytm.

[33] Roman Trentinaglia, Sven Merschjohann, Markus Fockel, and Hendrik Eikerling. Eliciting security requirements – an experience report. In Alessio Ferrari and Birgit Penzenstadler, editors, *Requirements Engineering: Foundation for Software Quality*, pages 351–365, Cham, 2023. Springer Nature Switzerland.

[34] Katja Tuma, Christian Sandberg, Urban Thorsson, Mathias Widman, Thomas Herpel, and Riccardo Scandariato. Finding security threats that matter: Two industrial case studies. *Journal of Systems and Software*, 179:111003, 2021.

[35] Tony UcedaVélez and Marco M Morana. *Risk Centric Threat Modeling: process for attack simulation and threat analysis*. John Wiley & Sons, 2015.

[36] Charles Weir, Ingolf Becker, and Lynne Blair. Incorporating software security: using developer workshops to engage product managers. *Empirical Software Engineering*, 28(2):21, Dec 2022.

What Motivates and Discourages Employees in Phishing Interventions: An Exploration of Expectancy-Value Theory

Xiaowei Chen[1], Sophie Doublet[1], Anastasia Sergeeva[1],
Gabriele Lenzini[1], Vincent Koenig[1], Verena Distler[2]
[1]University of Luxembourg
[2]University of the Bundeswehr Munich

Abstract

Organizations adopt a combination of measures to defend against phishing attacks that pass through technical filters. However, employees' engagement with these countermeasures often does not meet security experts' expectations. To explore what motivates and discourages employees from engaging with user-oriented phishing interventions, we conducted seven focus groups with 34 employees at a European university, applying the Expectancy-Value Theory. Our study revealed a spectrum of factors influencing employees' engagement. The perceived value of phishing interventions influences employees' participation. Although the expectation of mitigation and fear of consequences can motivate employees, lack of feedback and communication, worries, and privacy concerns discourage them from reporting phishing emails. We found that the expectancy-value framework provides a unique lens for explaining how organizational culture, social roles, and the influence of colleagues and supervisors foster proactive responses to phishing attacks. We documented a range of improvements proposed by employees to phishing interventions. Our findings underscore the importance of enhancing utility value, prioritizing positive user experiences, and nurturing employees' motivations to engage them with phishing interventions.

1 Introduction

Phishing was the most reported cybercrime in the U.S. between 2019 and 2022 [27]. Phishing emails deceive people into clicking on malicious links, disclosing sensitive infor-

mation, or installing malware on their devices [2]. Phishing attacks endanger organizational intellectual property and institutional reputation, causing billions of losses [4, 27, 40]. Organizations employ a range of measures to defend against phishing attacks. Despite the implementation of technical filters, even if deep learning models achieve an accuracy rate of more than 96% [7, 33], a substantial number of phishing emails still end up in employees' inboxes. While technical solutions play a critical role in mitigating phishing attacks, employees are the last line of defense in organizations [55].

To raise employees' security awareness and educate them about phishing attacks, some organizations deploy online security courses as a cost-effective way to educate their employees [18]. Some organizations utilize simulated phishing tests in an attempt to track whether employees can identify phishing emails [10, 22]. Further, organizations broadly advocate for employees to report phishing emails, which enables IT teams to promptly detect incoming phishing attacks [52]. Research suggests that phishing interventions promote safe responses to attacks [49, 81], and reporting can serve as an effective crowd-sourced approach to counteract phishing [12, 52]. However, these user-oriented phishing interventions are not always embraced by employees [51, 62], as participation in the interventions requires time and effort and can interrupt the working routine [31, 47].

Motivation theories from educational psychology can be useful in explaining employee's (dis-)engagement. Recently, Expectancy-Value Theory (EVT) has received attention from scholars working in information management [68]. EVT seeks to explain individual behaviors with two central constructs: "expectation of success" and "subjective task value" [14]. We find these constructs particularly relevant and underinvestigated in security behavior studies [15].

In this paper, we examine employees' engagement with *phishing awareness campaigns*, which include online security courses and simulated phishing tests, as well as *reporting phishing emails* through the lens of EVT. By deepening understanding of the influencing factors associated with phishing interventions, organizations can improve their implementation

of these interventions. We pursue the following objectives: 1) examining factors that motivate and discourage employees from engaging with phishing interventions, and 2) exploring what could be improved to increase employee engagement with these interventions. Focus groups are a qualitative method frequently applied to elicit ideas [79] and confront different viewpoints [78]. Educational institutions are frequently targeted by cybercriminals in recent years [53, 69]. Examining factors that influence university employees' engagement with phishing interventions is highly relevant to the current threat landscape. In light of this, we conducted seven focus groups with 34 employees (including research and non-research roles) in a European university.

Contributions. This paper makes empirical contributions, providing an enriched understanding of how various factors influence employee (dis-)engagement with phishing interventions. Our findings and adaptation of EVT suggest that it is a valuable theoretical framework for explaining how motivational factors influence employees' engagement with phishing interventions, highlighting its potential as a framework for future security behavior studies. This paper makes a theoretical contribution and highlights the possible adaptations to EVT for future use in organizational cybersecurity. Additionally, we offer practical suggestions for improving phishing awareness campaigns and reporting procedures in organizations, advocating user-centric approaches.

2   Related work

2.1   Phishing awareness campaigns

Simulated phishing tests are a tool for both assessment and educational purposes at organizations [22, 39]. Prior studies primarily utilized employees' click-through and reporting rates in phishing tests as indicators of employees' security behavior and their resilience to phishing attacks [22, 49, 81]. A recent case study highlighted that conducting simulated phishing tests at an organization requires significant time and effort from different stakeholders [8]. Moreover, some organizations have experienced side effects from phishing tests that have burdened CISO's relationship with employees [39]. When organizations neglect privacy concerns, fail to receive approval of simulated materials, don't specify the purpose of tests, or withhold appropriate feedback, it can lead to negative reactions from employees [62]. Phishing tests also increase employees' workload, potentially making them more susceptible to phishing attacks [8, 62]. Brunken et al. suggest involving employees in future research to better understand how simulated phishing tests impact them and their overall productivity at the workplace [8].

A variety of formats have been introduced to engage individuals with online security training [42]. Comic and game-based online trainings have reported notably high levels of

satisfaction in user evaluations [50, 75]. A meta-analysis revealed that trainings combining text and comics demonstrated large effects in reducing victimization compared to comics or game-based trainings [9]. Online phishing quizzes, such as jigsaw puzzles, effectively improved participants' skills in detecting phishing emails [74]. Volkamer et al. created and evaluated a five-minute phishing awareness video, which significantly enhanced participants' ability to recognize phishing attempts both immediately and after an eight-week interval [72]. User feedback praised their video's clarity and simplicity, with suggestions for more phishing examples and a concluding summary [72]. Anti-phishing training utilizing storytelling led to higher levels of curiosity, self-efficacy and phishing detection ability than training employing comics in an online experiment [43]. To improve the effectiveness of security trainings, both the content and format of trainings were re-designed to engage learners.

While some studies suggest that offering educational materials after simulated tests improved employees' safe responses to phishing [49, 81], there are concerns about the effectiveness of this embedded training approach [8, 52]. Kumaraguru et al. found that employees who trained with anti-phishing materials after clicking links in simulated phishing emails exhibited a decreased likelihood of clicking on links in subsequent phishing tests compared to their untrained colleagues [49]. Yeoh et al. reported that the immediate provision of anti-phishing materials following phishing tests led to more safe responses than merely administering phishing tests [81]. Despite these findings, researchers suggested that only a small percentage of employees who clicked the phishing tests subsequently engaged with training materials [8, 20]. Thus, further investigation is required to better integrate simulated phishing tests and online security courses.

2.2   Phishing email reporting

Recent studies have begun to investigate factors that influence individuals' intention to report phishing emails. A survey with American college students [51] revealed that perceived self-efficacy, expected negative outcomes (*concern for mishandling of reports of spear phishing emails*), and cybersecurity self-monitoring increase the likelihood of reporting spear phishing emails. In alignment with [51], Kersten et al. suggested that user's intention to report phishing emails was negatively associated with the perceived "believability of the email" (the extent a user considers the email to be credible) in an online controlled experiment [46]. In an in-situ deception study [20], Distler found that employees' motivations for reporting phishing included improving email filters and receiving positive feedback. Obstacles to reporting entailed uncertainties regarding the reporting process and rationale, coupled with concerns about "getting colleagues into trouble" for sending legitimate emails that were misinterpreted as phishing attempts. Additionally, participants believed that

reporting became redundant once they had clicked on the link in a simulated phishing email [20]. In a survey with US workers, factors such as self-efficacy, subjective norms, and altruism tendencies increased reporting intention. Conversely, "sportsmanship" hinders individuals from reporting phishing emails [56]. Other than utilitarian motives, Franz proposed that the design features and risk indication influence participants' acceptance of reporting tools and suggested further research into the role of hedonic motives in the reporting process [30]. Additional factors may influence an individual's intention and behavior regarding the reporting of phishing emails, warranting further investigation.

2.3 Theoretical models applied to study user security behavior

Prior research on user security behaviors has frequently focused on fear appeals, as seen in studies that examine the constructs of Protection Motivation Theory (PMT) [36]. PMT explains protection behavior through two processes: threat appraisal and coping appraisal. In threat appraisal, people evaluate their perceived vulnerability and the perceived severity of a situation, while coping appraisal entails assessing response efficacy, self-efficacy, and response cost [58, 63, 70]. However, there are limitations and constraints in applying PMT to study user security behaviors. Originally constructed to explain health protection behaviors, PMT is based on the premise that the threat is relevant to the individual; however, this might not be the case in the information security context [57]. In a Relative Weight Analysis, attitude, personal norms & ethics, and normative beliefs demonstrated the highest effect sizes and relative importance in explaining security compliance behaviors, emphasizing employee psychological and ethical traits [15]. These constructs are not included in the theoretical model of PMT. To overcome the limitations of PMT, recent studies have begun to integrate constructs from other motivational theories to examine user security behaviors [36].

Expectancy-Value Theory (EVT) [23] is an influential motivation theory in educational psychology [38]. According to EVT, individuals' beliefs about how well they will do on an upcoming task and the subjective values they attributed to it influence their engagement with the task [25] (refer to Appendix A for the core constructs of EVT). EVT shares the same theoretical root as PMT, as both theories developed from Atkinson's expectancy-value model [63, 76]. EVT examines individuals' anticipation and subjective task values in educational contexts [76], whereas PMT employs fear appeals to motivate protective actions in health management [63]. However, EVT has rarely been applied to security behavior studies [15]. In an experiment incorporating EVT constructs, Jenkins et al. found that the highest levels of security behavior were associated with minimal technical controls (*number of passwords a participant was forced to use and remember*)

combined with security education [44]. A recent structural modeling study that applied EVT revealed that achievement, along with intrinsic and extrinsic motivations, are determinants in explaining the motivational values associated with users' intention to protect mobile identity [3]. Applying EVT to investigate the factors that influence employee engagement with phishing interventions appears promising.

2.4 Research objectives

Low employee engagement with phishing interventions continues to be an obstacle to achieving information security in organizations [51, 81]. EVT has been utilized to examine learners' motivations in various contexts, including organizational [11, 41]. Applying EVT can elicit employees' motivational factors associated with phishing interventions. Further, beliefs and values form attitudes in the cognitive process, which in turn guide behavioral responses [44]. Expectation and subjective task values directly influence people's choices and performance in the EVT framework (see figure 1). Consequently, we propose to utilize EVT constructs to address the following research questions (**RQ**):

> RQ1: Which factors motivate employees to engage with phishing interventions?
>
> RQ2: Which factors discourage employees from engaging with phishing interventions?
>
> RQ3: From the employees' perspective, which aspects of phishing interventions could be improved?

3 Study design

We conducted focus groups with 34 employees at a European university to address these research questions. Focus groups are a form of interviewing where multiple participants come together to express and deliberate on their views regarding a predetermined topic in a collective discussion [21]. Focus groups are especially useful for gathering diverse and in-depth perspectives from interactions among participants [32, 78, 79], allowing us to gain an exhaustive understanding of the factors influencing employees' engagement with phishing interventions.

3.1 Study context

The study was conducted at a research-oriented European university that employs approximately 3,900 individuals. 38% of them are employed in research roles, whereas the remaining employees fulfill administrative functions. The organization uses a phishing awareness campaign sourced from a security service company. The IT team sends a simulated phishing test to all employees via the management software on a random

date each month. Employees who click the link or download the attachment within the phishing test land on a page displaying "you clicked on a simulated phishing test" and "rules to stay safe online". Afterwards, the IT team sends a web link to online security courses to those who responded unsafely. Employees who reported the simulated email to the IT team receive an automatic reply within a couple of minutes with the subject line "congratulations, you've spotted a phish".

To raise phishing awareness, the IT team sends every new employee an email during their first week that includes links to online security courses and suggested responses to suspicious emails. To defend the organization against phishing attacks, the IT team encourages employees to report any suspicious emails to "report-a-phish@anonymized". When the reported email is a simulated test, a program automatically sends out a reply; otherwise, a security expert manually reviews the reported email. Normally, it takes one or two working days for the expert to reply with the verification result of the reported email. When a reported email is a phishing attempt, the expert sends a phish alert to individuals who also received the phishing attempt. When the email is legitimate (not a phish), the expert replies with "It is a legitimate email".

At the time of our investigation, all employees automatically received simulated phishing emails as part of their cybersecurity training without prior informed consent. Employees could either actively engage by reporting the simulated test in accordance with the organization's suggestions for handling suspicious emails or ignore these simulated tests.

3.2 Participants

We used multiple approaches to recruit study participants, including posters across three administrative buildings, LinkedIn posts, email invitations, and direct outreach. Forty-five employees registered their interest in participating in our study. We assigned them to different groups based on the similarity of their job roles and the diversity of faculty. We did not exclude any specializations (e.g., computer scientists) when scheduling our focus groups. Due to personal reasons, 34 of the 45 interested employees participated in seven focus group sessions (20 female, 13 male, and one non-binary) between November 2022 and January 2023. Each session consisted of three to seven participants. Participants included 19 researchers, 12 administrative staff, and 3 software developers. On average, the research staff had worked at the organization for 1.3 years (SD=0.9), and the non-research staff 7.3 years (SD=6.7). The participants' age ranged from 25 to 56 years (mean=37.6, SD=10.8). In the demographic questionnaire, 32 (94%) participants indicated that they had encountered phishing attacks previously; 29 (85%) had received simulated tests from the IT team[1]; 25 (74%) had reported phishing emails

---
[1]Every employee is scheduled to receive a phishing test monthly. These five employees, who reported not receiving any phishing tests, may have simply not clicked on or noticed the tests.

to the IT department, and 14 (41%) had previously participated in online security courses. We include the participant demographic information in Appendix D.

3.3 Procedure

Prior to data collection, we conducted two pre-test sessions (N=11) to refine our protocol. During the first pre-test, we led the discussion using a synthesized framework of motivation theories [38]. Introducing concepts from multiple theories led to cognitive overload for participants during the focus group. In the second pre-test, we narrowed our focus to EVT. According to the preliminary analysis, observations, and participants' feedback on the pretests, we improved our discussion questions and added templates and brainstorming activities. The revised focus groups included four parts: a warm-up activity, a group discussion, a brainstorming activity, and the debriefing. Each focus group took approximately 90 minutes.

First, we conducted a warm-up activity to familiarize the participants with the lab and to elicit what motivates and discourages them from engaging with a self-selected leisure activity through **Template 1**. This stage lasted for 10 minutes.

In the second part, the participants were involved in a group discussion on phishing awareness campaigns for 25 minutes. Then, we instructed them to complete **Template 2** to record their motivating and discouraging factors for reporting suspicious emails. Following this, participants continued discussing the factors influencing their reporting. This stage planned a total of 60 minutes and included 12 questions to examine *general opinions, self-concept of their ability, goal setting*, and *role identification*, as well as their subjective task value (*costs, benefits*) related to participating in phishing interventions. These questions were adapted from the core concepts within EVT framework that affect individual's choices and performance (see Figure 1).

In the third part, participants were asked to brainstorm as if they were the new chief information security officer in response to an increase in phishing emails targeting the university. Participants were tasked with designing strategies to engage employees with phishing interventions in groups. This round lasted 15 minutes.

Lastly, the participants were debriefed by introducing the standard practices suggested by the IT department to avoid any misunderstandings caused by opinions mentioned during the discussion. We provide the **two templates** and full focus group **protocol** in Appendix B.

3.4 Data collection and analysis methods

We recorded audio and video of the focus group sessions. We used the audio recordings (11 hours in total) for the analysis[2]. The audio was transcribed automatically using Microsoft

---
[2]Videos were recorded with the lab's default system as a backup resource in case of audio disruption and were deleted after transcription.

Word and reviewed to ensure accuracy. We pseudonymized the transcripts to protect the identity of participants prior to analysis.

The answers to "Template 2. What motivates/discourages you from reporting" were transcribed into an Excel spreadsheet. The first and second author then independently coded the template, following a thematic analysis procedure [13]. Then the two authors categorized the generated codes into preliminary groups in a discussion, which yielded an initial set of codes. Concurrently, a coding workshop was conducted with five researchers experienced in qualitative research and coding. This workshop, which employed an inductive approach [34], analyzed the transcripts from two focus group sessions. Consequently, a second set of codes was created. By integrating the template codes with those from the workshop, the first author established a code system in MAXQDA [71]. The code system was reviewed and revised by three authors. All transcripts were subsequently coded by the first author using MAXQDA. Theme saturation [59] was reached after completing the coding of data from the sixth group. The second author thoroughly reviewed all coded transcripts for consistency and accuracy. A few disagreements were resolved before the final summary of findings via discussion between authors and reviewing the context of the coded segments. We include our coding scheme in Appendix C.

3.5 Ethics

The study received approval from the university's ethics review board prior to the pretest. We emphasized that "the session is strictly confidential" to assert peer confidentiality in the email confirmation prior to each session. All participants were informed of their right to withdraw both during and after the study and provided informed consent. The raw data collected in this study were kept confidential to the researchers and stored in line with the General Data Protection Regulation (GDPR) and the ethical guidance of the research institution. Each participant received a €40 gift voucher as compensation for their 90-minute participation. We only used pseudonymized data for analysis.

4 Results

We present the factors thematically according to the core concepts of EVT framework and highlight those that could not be located within the framework (see Table 1). Unlike qualitative data from individual interviews and open-ended questionnaires, the factors emerging from focus group conversations represent a co-creation among participants. There were occasions when participants filled in specific factors in the template (e.g., P28: "being a good citizen") but did not mention them during discussions, or situations where a factor was articulated in depth by one participant, leading others to

choose not to repeat it. Providing the frequency of each theme mentioned by participants would thus not be meaningful.

4.1 Phishing awareness campaigns

4.1.1 Factors that motivate employees

Gaining phishing knowledge and *enhancing phishing awareness* are the two utility values mentioned by many participants. They noted that the awareness campaign demonstrated that phishing attacks are constantly changing and evolving. They learned that it is critical to remain informed of evolving phishing techniques, which can support their decision-making in responding to suspicious emails. Additionally, phishing campaigns keep them vigilant of phishing attempts in their daily work. Not only beginners who were not tech savvy could benefit from the campaigns but also experienced employees could be reminded that they need to be cautious of contextual factors. As P2 stated, "even if you're aware of the problem and know how to check . . . you can still fall for it (phishing test) if you don't pay attention, if there's a lot of stress and you're going faster." Additionally, a few participants considered participating in phishing campaign to be a game (P8), and some parts of the online training were "awesome" and "*fun*" (P26).

Acquiring skills in identifying whether emails are legitimate or not from awareness campaigns was mentioned by some participants as a motivating factor. Through the campaigns, they increase their competence (self-concept of one's ability). They perceived the phishing campaign as beneficial in "training people to recognize what is phishing and prevent them from actually falling into one when it happens" (P22). Consequently, they held this expectation of maintaining *cyber safety*. As P9 shared, the campaign not only benefited them in terms of protecting their own data and e-mail accounts, it also "helped the university as an institution to be better protected."

A few participants believed that receiving training on security-related knowledge could benefit their life and improve their computer literacy, contributing to *personal development* or long-term goals. P29 stressed that cybersecurity knowledge would become a fundamental skill for them to perform daily tasks with digital tools, and "it's not only about fear of being attacked, you need to understand what's inside these technology tools ... everything related to cybersecurity is very fundamental now and, in the future, would become even more fundamental, like reading."

4.1.2 Factors that discourage employees

Perceived low value discourages participants from taking online security courses, as indicated by P9, "not sure this kind of course will help me to be more precise in making judgments." On the one hand, the course was perceived as low value for some participants who had received security training before working in the current organization. On the other hand, some

Table 1: Motivational factors associated with phishing interventions.

	Phishing Awareness Campaigns		Report Phishing Emails	
	Motivating	**Discouraging**	**Motivating**	**Discouraging**
Expectation	Cyber safety	Optimism bias	Expectation of mitigating, Fear of consequences	Lack of feedback, Lack of communication
Utility value	Phishing knowledge, Phishing awareness	Perceived low value, Lack of incentive	Protecting oneself, Safeguarding the workplace	Low utility value
Intrinsic value	Fun	Lack of interest	Enjoyment, Satisfaction, Pride	
Attainment value		Other priorities	Core values	
Cost		Time constraint, Interrupting workflow, Opportunity cost, Negative inference	Easy to report	Usability issues, Worries and privacy concerns
Competence	Acquiring skills	Overconfidence	Empowerment	Low self-efficacy
Social identity			Recognition, peer influence, sense of belonging	
Goal	Personal development			
Self-schemata		Procrastination		Habitual behavior
Previous experience		Fear of failing the training	Phishing experience	
Outside of EVT				Contextual factors

participants had concerns that the course might be in technical language, which can be difficult for people who are not tech-savvy to understand, "I'm going to attend it, but I'm not going to understand it" (P13). Furthermore, participants shared that the *lack of incentives* discouraged them from participating in security course. If the organization offered incentives, such as course credits (for doctoral researchers), compensation, and praise from the team leader, they would be more likely to participate in the security courses. As P24 asked, "what is my incentive to do an optional course here?"

Some participants expressed that even though they had intended to learn from the security course, the cover image and name of the course gave them the impression that it would *not be interesting*, resulting in them disengaging with the courses (P16). Participants thought that the course exercises were too simple; "the exercises were so obvious that you would truly have to make an effort to answer wrongly" (P2).

Participants frequently mentioned *time as a constraint* that discourages them from engaging with awareness campaigns. Participants found it difficult to allocate time to the awareness campaign due to their packed schedules. Time spent on the campaign was seen as an *opportunity cost*, as P23 stated, "instead of achieving something for your project, for example, a good experimental result, you spend time on the phishing campaigns, and you lose that opportunity." Multiple participants shared that a downside to engaging with awareness campaigns was heightened worry about potential threats - "*Negative inference*" (P30). An awareness campaign might lead them to experience more stress, compelling them to exercise increased caution in their daily lives (P5 and P25).

Participants expressed less interest in the campaign if the course content was not relevant to their area of expertise or interests. "*Other higher priorities*, such as course work and the experiment, would discourage me from participating in the awareness course; for me, the security courses were super boring" (P23). Participating in awareness campaigns requires people to switch from their tasks at hand to phishing-related content. The switching *interrupted their workflow* (P25). Switching between tasks meant that it took additional hours for them to perform their duties (P27).

Participants' belief that they were less likely to experience phishing compared to others led to less involvement with the awareness campaign (*optimism bias*). As illustrated in P14's case, "I always had this thinking, it won't happen to me because this (phishing email) is so stupid." Participants also indicated that *overconfidence* in their knowledge of the topic

made them less likely to engage with the awareness campaign (P28).

Previous negative experiences with security courses might evoke a *fear of failing the training*, which discouraged employees from participating. As P8 shared, "the fear or the worry that if I failed the course, it would be tracked. Because I experienced that in the previous job. If you didn't get a certain grade, then you would be forced to retake it and retake it." Additionally, participants shared that *procrastination* resulted in delaying or forgetting to take the courses (P32 and P33).

4.2 Report phishing emails

4.2.1 Factors that motivate employees

Participants had specific *expectations* when they reported phishing emails. Reporting was a practical way of notifying colleagues and alerting them of phishing attempts. Participants expected that the organization would improve its spam filters with their reported emails, which would benefit them in terms of receiving fewer spam and phishing emails in future. "The main benefit of reporting is that the IT team could create more filters for phishing emails if they have more data (from reporting), making us safer" (P27). They expected that the organization could contain the damage, retrieve stolen data from attackers and mitigate risks. *Worries and fears* related to the consequences of phishing attacks prompt participants to report. Specifically, participants worried that they would get into trouble, lose information, suffer from financial risks, and involvement in cyber crimes if they did not report promptly. Several participants emphasized reporting to avoid potential reputational damage and financial losses for their workplace (P13).

Participants indicated that reporting *protected their personal data*, financial assets, and other valuable possessions, including personal accounts. When suspicious of an email, they received support from the IT department in assessing the reported email. Beyond work-related protection, one participant felt safer in their personal life after reporting a phishing attempt to law enforcement, specifically an email accusing them of financial misconduct. Their concerns were alleviated once the email was confirmed as a phishing attempt. Participants also regarded reporting as a measure to *safeguard the workplace*. Firstly, reporting phishing attempts protected the organization's confidential data, documentation, work tools, internal network and servers from external access (P23). Secondly, reporting was viewed as a way of raising awareness of phishing attempts in the organization. Not only the IT team needed to be notified of phishing attempts, but also their colleagues (P11 and P12). Thirdly, participants regarded reporting as a collaborative approach to countering phishing. The IT team assisted the employees in verifying the legitimacy of emails, and employees assisted the IT team in detecting the phishing attempts in real-time (P19).

Participants shared their *experiences receiving phishing emails*. Some received suspicious emails from professors, colleagues or family members asking for money or directing them to fraudulent websites. Others fell for phishing attempts while using online hotel booking platforms. P19 is a doctoral researcher in computer science who got phished a week before the focus group, "I lost two days of my life trying to correct just one click. During the backup, I lost a bunch of documents (erased a password for storing work documents), so there were other consequences after that." Even though the incident happened in their private life, it impacted their work. After the phishing incident, P19 wanted to warn others about phishing attacks and was motivated to report phishing attempts.

The ease of reporting phishing emails was mentioned as a reason why some participants reported phishing frequently. They referred to the one-click reporting button as straightforward, which made the reporting process simple and not time-consuming. They emphasize the one-click option for quick responses. The positive user experience of the reporting button facilitated participants to report, as exemplified by P31: "It's easy so it doesn't take even two seconds. If you suspect, click, click, and then you're done."

Participants regard the "congratulations" email that they received from the IT team when they reported a (simulated) phish as a kind of "*recognition*" and extrinsic reward for their reporting (P9). While P21 used to ignore phishing emails, one colleague told them it's better to report (*peer influence*). After that, P21 started to report suspicious emails. The *sense of being part of the community* prompts participants to report, as exemplified by the following conversation:

> P32: "We need to participate. We're all active users and it's not just IT who has to deal with it."
>
> P34: "We are actors within the community. So, we are together."

Participants described that they experienced feelings of *enjoyment, satisfaction*, and *pride* when reporting phishing attempts, likening the process to a game, feeling proud of their vigilance, and deriving a sense of satisfaction from reporting. As P28, P11, and P8 indicated:

> "When you click to report phishing attempts, then you receive 'congratulations'. I'm happy and it's like a game." (P28)
>
> "I can relate to the sense of satisfaction. Once you've reported it, you feel like you played your role. You did a good job." (P11)
>
> "I don't want to break my streak of always reporting the phishing attacks ... I'm quite proud of that." (P8)

Several participants mentioned a number of *core values* (guiding principles that shape people's attitudes, actions, and decisions) that drive them to report phishing attempts, including "help others" and "vulnerable" groups (P2 and P15), "duty"

(P11), "being a good citizen" (P28 and P33), and "contributing to the fight against phishing" (P33). Additionally, a few participants considered reporting as an approach to take control and make a difference (P6). In P16's case, "I had the initiative to defend against the phishing attack. And knowing that I can stop spreading this attack for other people and for my future self really helps me, like *empowering*."

4.2.2   Factors that discourage employees

Multiple participants felt discouraged from reporting suspicious emails because they received *no feedback* on the outcome of their actions. They expected to receive more information about the outcomes of their reporting (P12). As P31 emphasized, "we don't know what the effectiveness of reporting phishing emails is. We don't know the numbers, so it would be really good to have a kind of feedback status. What has been done last year? What was the success rate?" Further, even for participants who reported diligently, they sometimes felt discouraged from reporting due to not knowing whether their colleagues were reporting or not (*lack of communication*).

> "I report phishing emails regularly and religiously, but I'm thinking is everyone else doing the same as me, putting in the same effort as I am on reporting? It takes maybe 30 seconds of your time, but I'm still very careful about it." (P25)

The perceived *low utility value* discouraged participants from reporting phishing emails. Firstly, the belief that the "phishing" email is merely a test from the IT department reduces the perceived need to report it, as stressed by P27, "for me, every phishing email that I received was a simulated one. So, I didn't see the point of reporting that because I knew that it was from IT." Secondly, if the participants believed most people would be able to recognize the email as a phish and posited a low threat to others, they chose not to report (P16). Thirdly, worries of additional burden due to reporting discouraged participants from following the reporting procedure. These assumed negative outcomes included "bog me down with questions" (P13), getting "more emails" (P17), and "fear of annoying IT staff" (P28). Lastly, the belief that reporting doesn't lead to effective outcomes, such as prevention or resolution of the attack, discouraged participants from reporting. As exemplified by P19, "the lack of results discourages me. It seems like we try to do something nice and nobody really cares."

Participants highlighted several issues related to ease of use, functionality, and efficiency in the reporting process as discouraging factors (*usability issues*). Some participants found the reporting procedures ambiguous. For instance, P8 only learned about the "report-a-phish" email address from a colleague after observing the absence of a reporting button following an update of the email client. P26 wondered about the preferred method of reporting, stating, "I forwarded it to report-a-phish, and they said, 'Oh no, can you please send it as an attachment instead of forwarding it'." For participants who frequently reported suspicious emails on their laptops mentioned that they often delete or disregard such emails when viewed on their smartphones. P9 shared, "I wanted to report it and I had trouble doing that with my phone. So I always try to be extremely careful, almost like you have something burning in your hand." Despite their caution, they still accidentally clicked on the email when trying to report it, leading them to ignore phishing emails on their phones. Moreover, Linux and Mac OS users felt the reporting process demanded too much effort. It's easier to just delete the suspicious email than to forward the email as an attachment to the IT department. As emphasized by P24, "if it's anything more than a one-button click would be a little bit more discouraging."

Participants expressed they would not report when they were concerned that the suspicious emails "*disclose their private information*" or cause false impressions about their personal life (P4, P28). Additionally, *worries about being judged* by the IT team were shared as a discouraging factor by participants. As the conversation between P33 and P34 revealed:

> P34: "I have this feeling that IT guys, they're always like a bit, 'they don't know they're doing really.' And I feel I'm so stupid. If I report Netflix or something as phishing, then they would think 'stupid woman'."
>
> P33: "They could judge us."
>
> P34: "So this feeling unnerved me and discouraged me from reporting. Because they give you this feeling sometimes. I experience it, I call the help desk and get this 'again'."

Participants shared that they frequently postponed or forgot to report because they reverted to their *old habits* of simply deleting emails. They mentioned that the reporting process is unique to their current workplace, contrasting it with their usual habit of deleting or marking suspicious emails as spam. As P11 stated, "in my personal life, when I encounter a suspicious email, I just delete or mark it as spam. However, this report-a-phish button is quite specific and new." Participants noted that if they *lacked confidence* in identifying whether an email is phishing, they would typically ignore it. Furthermore, some participants cited "laziness" as a reason for not reporting.

Contextual factors, such as task overload, stress, and time pressure, could deter participants from reporting phishing emails. When focused on one's tasks and in the status of flow, they perceived incoming emails as a distraction, resulting in less intention to report (P27).

4.3 Improvements proposed by participants

Participants proposed various ideas to make phishing interventions more engaging during the brainstorming sessions. We categorized them into the following themes:

Gamification elements: Participants suggested adding achievement, competition, virtual reputation, and fun elements to the reporting process. There should be rewards or acknowledgments for the department that actively participates in awareness campaigns and reports the most phishing emails. Participants recommended providing incentives for participation in phishing campaigns, such as gifts, praise, and course credits. Participants suggested that role-playing and leaderboards would engage employees with the security training.

New employees & Mandatory training: During the onboarding week for new employees, the university should provide a mandatory training session to equip them with knowledge about phishing and the reporting procedure. The IT team should walk in the shoes of new employees and find out the potential attack points within their work activities. Participants also suggested making a security course mandatory for frequent clickers of phishing tests and for departments that receive a high number of phishing attacks.

User experience: Participants suggested to improve the user experience of phishing interventions. Real-time verification of reported emails and shorter, more relevant and interactive trainings would attract employees. Course content should be personalized according to different levels of phishing knowledge. Participants suggested using pop-up quizzes instead of online videos to raise phishing awareness because the latter took too much time.

Communication: Participants suggested that the IT team provide regular updates or host information sessions with employees. The positive impacts that phishing interventions have on the university should be communicated quarterly or annually. Seminars drawing from diverse expertise areas like IT, HR, and research were recommended to bolster organizational defense and collaborations between departments.

Feedback: The IT team should gather feedback on phishing interventions from employees, provide statistics on phishing interventions, and be transparent about the state of the art and the efficacy of current solutions. Participants also suggested the IT team provide individual feedback on what happens after an employee reports phishing.

Present real incidents: Participants suggested the IT team present real phishing attacks and their consequences as examples to raise awareness. Providing concrete examples of how data breaches happened through phishing would raise employees' phishing awareness.

Authentication of internal emails: Participants suggested implementing digital signatures to authenticate internal communication, which would enable fast detection of phishing emails that masquerade as internal communication. Additionally, participants suggested recruiting *more IT employees* to host training sessions regularly, noting that the IT team seems occupied with an overload of tasks. Lastly, one group proposed a *punishment* approach, that is, increasing the number of simulated phishing emails for employees who repeatedly clicked simulated phishing emails.

5 Discussion

5.1 Applying EVT to the context of organizational cybersecurity behaviors

In this study, we investigate how Expectancy-Value Theory (EVT) can illuminate the factors influencing employees' engagement with phishing interventions. Building on our findings and considering that EVT was created to interpret achievement-related choice and performance in educational settings, we propose incorporating an organizational dimension into EVT model (refer to Figure 1, our adaptations to EVT are in blue italics). Hence, we suggest integrating the organizational dimension in the form of "organizational culture" [80] into a "cultural milieu" construct, which can be described as a system of social roles, each with its associated responsibilities and obligations [77]. Perception of the organizational dimension can be interpreted through the lens of the "psychological contract", which refers to an unwritten set of expectations and beliefs about the obligations that exist between an employee and their employer [35], also including employees' beliefs about their responsibilities in organizational security [37]. During group discussions, employees consistently highlighted that, through their security behaviors, they aim to collaborate with the IT department in fighting against phishing attacks and safeguarding the organization. Despite the absence of explicit organizational policies dictating such obligations, this inclination can be attributed to the implicit norms acquired through the organization's unspoken rules and in general - organizational culture as a proxy for information security culture in the organization [66]. Our results suggest that the perception of the organizational culture, communicated through socializers' beliefs and behaviors, can contribute to a constructive "us vs. them" (organization vs. attackers) mentality, where employees have a self-concept of a contributor to organizational security.

In accord with past studies [19, 73], we observe that "peer influence" and "knowledge sharing" among colleagues influence employees' intention to report phishing emails and participate in online security courses. Pursuing this line of thought, we can extend the EVT model's "socializer" construct to include "colleagues and supervisors." These people convey their knowledge of the organization's unwritten norms to other employees, aiding in shaping security protective identities. Furthermore, we propose that employees' security consciousness stems from their social identity in EVT. Being a "responsible" employee dedicated to the organization, in harmony with other foundational roles, makes up one's social
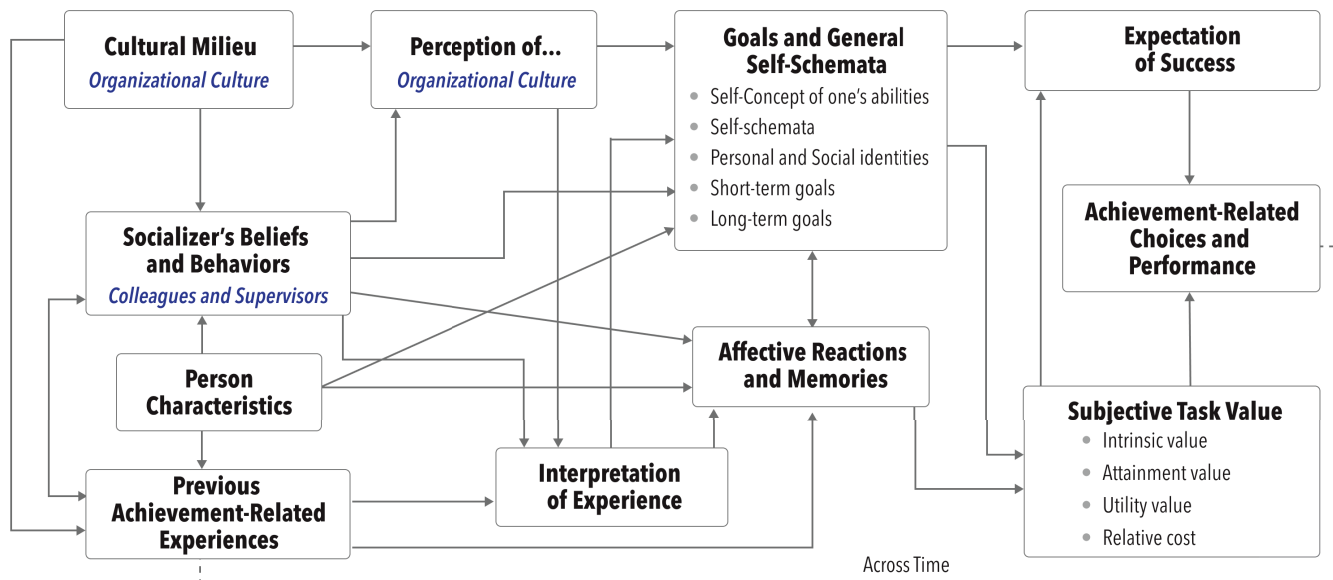
Figure 1: The expectancy-value model (adapted from [25]).

identity. This is connected with the "extra-role security behaviors" phenomenon [29, 54], in which some employees are self-motivated to take additional responsibilities to secure the organization, even if these responsibilities go beyond their contractual role. In our study, we found strong evidence that this type of motivation is one of the core drivers of reporting phishing emails.

In summary, our empirical findings demonstrate that the EVT framework can be specialized for use in organizational security settings. We specify certain concepts of EVT in the organizational setting, proposing to focus on the organizational culture, colleagues and supervisors, and "previous experience" on the left side of the framework (see Figure 1). Our findings also support the original EVT framework with findings that subjective task values, expectations, goals, and general self-schemata influence employees' engagement with phishing interventions. The discovery paves the foundation for future studies to apply EVT in studying organizational security behaviors.

5.2 Subjective task value of phishing interventions

The majority of educational interventions based on EVT focus on altering individuals' "Subjective task value" [26]. Subjective task value is the core construct within EVT, in which the value of engaging with an activity can be considered as the ratio between perceived benefits and associated costs [25]. People tend to opt for activities that have a higher benefit-to-cost ratio. Our findings showed that many of the discouraging factors of phishing awareness campaigns are associated with

different types of costs, such as psychological cost, time cost, and opportunity cost. The findings align with previous studies on imposing security measures within organizations, which found that employees perceived the security measures as extra burdens that encumber their work [8, 47, 62]. Previous literature proposes remedies such as reducing the friction associated with security measures and automating security protocols [16, 31].

In the EVT framework, another promising avenue for exploration emerges: the potential for security managers to tip the scale in favor of security measures by reducing their associated costs. This shift could engage employees more with security measures. This idea aligns with studies showing positive outcomes from security trainings in short video format, with participants regarding the training "informative"/"useful" [82] and expressing interest in extended sessions [72]. The increased benefit-to-cost ratio in such cases can be attributed, in part, to the brevity and density of the training content. Our study echoes employee preferences for succinct training, as exemplified by "don't give me a half hour course for two minutes' value" (P13) in the group discussion. Similarly, participants in different groups proposed providing employees with shorter but more frequent security trainings.

Our study identifies a cluster of motivators associated with the intrinsic values of reporting. These motivators, deeply embedded in employees' psychological needs and desires, include satisfaction, empowerment, and core values (citizenship and altruism). Our findings are congruent with previous studies, which suggest that autonomy, personal values and principles influence users' security behavior [48, 57]. These elements, often sidestepped in security behavior research,

weave a complex network of factors influencing phishing reporting intentions. Considering that security messages that appeal to individuals' desires are more likely to elicit secure responses than those based on fear [57, 67], organizations should establish reporting procedures that resonate with employees' psychological needs. Integrating "fun" [65] and "experiential learning" [12] elements into training programs can enhance their intrinsic value, thereby engaging employees with phishing awareness campaigns. Furthermore, Eccles and Wigfield suggested developing attainment value-based interventions [26]. These interventions could take the approach of informing employees about the connection between anti-phishing practices and their personal values.

5.3 Previous experience, expectation of success, and personal development

Our study reveals that even motivated employees can become disheartened if they lack clear feedback and perceive their actions as ineffective. Several discouraging factors for phishing interventions can be categorized under "previous achievement-related experience." According to EVT, the "interpretation of experience" can influence "expectation of success" by altering goals and subjective task value. This attenuation is often due to negative experiences from prior engagement with the task. Employees are more inclined to adhere to security protocols if they deem the processes effective in mitigating phishing attacks [65]. Various employees in our study identified the lack of feedback and clarity about subsequent steps after reporting an email as discouraging factors, often provoking uncertainty and negative emotions. Such a phenomenon was also observed in employees' attitudes towards phishing awareness campaigns where previous unfavorable experiences shaped their perceptions. Over the last 20 years, research has persistently emphasized the critical role of feedback in fostering secure behavior within organizations [1, 6, 64]. Our study further explores the mechanisms through which an absence of feedback can alter motivation, even for motivated employees.

Intriguingly, we noted that prior experiences with being phished emerged as a strong motivator for some employees to report phishing, propelling their goal to prevent others from undergoing similar negative consequences. We hypothesize that the negative experience altered the subjective value they placed on reporting, which necessitates further study of this transformation from victim to defender in the context of phishing. Recognizing this transformative process can inform the development of support structures within the workplace. Employees who encounter cybersecurity incidents often experience guilt and shame. Workplaces should provide support, instead of blaming, to contain damage caused by the incidents and empower their employees [20, 61].

Employees demonstrated interest in acquiring security-related knowledge, linking it with their personal and professional growth. This interest suggests a pathway for organizations to refashion their security training to better align with employees' long-term goals. Given that all employees manage valuable accounts and passwords, and are often influenced by media reports or personal experiences of cybersecurity incidents, the imperative to adeptly navigate digital protection is clear. Similarly, Reeves et al. suggest shifting from a compliance-driven to a user-driven approach in security training to enhance the efficacy of training programs [60]. Incorporating employees' personal learning needs into organizational training paradigms could motivate employees to engage with security trainings.

5.4 Practical implications

We found that many of the discouraging factors related to the phishing awareness campaign are associated with its perceived value. Several usability-related factors discourage employees from reporting phishing emails. Fear, worries, and concerns about phishing interventions discourage employees from engaging (see Table 1). Leveraging insights from both the employee-generated suggestions and the EVT framework, we have proposed several improvements:

For phishing awareness campaigns: Clear communication of the campaign guidelines, expectations, goals, and consequences can alleviate the discouraging factor of "fear of failing training." Specific time slots should be allocated for employees to participate in the training sessions, addressing the discouraging factors of time constraints and interruption to their workflow [79]. This might not be possible in the case of knowledge workers who autonomously allocate time and tasks, for whom training will inevitably cut into their "productive" time. Making the training content relevant to individual job roles would enhance its relevance and applicability to daily tasks. Regular updates on evolving phishing attacks should be provided to increase awareness among employees. Gamification elements in the training program might enhance engagement [65].

For reporting phishing emails: Organizations should clearly communicate how reported incidents are managed by the IT team [28]. Timely feedback mechanisms should be established, reinforcing employees' sense of contributing to security. Regular updates (e.g., intranet, messages, displays) are beneficial for keeping employees informed about security efforts and emerging threats. Providing statistics on reporting and organizational benefits can underscore the personal value of reporting incidents. The reporting process should be frictionless to alleviate usability concerns. Ongoing awareness initiatives can foster engagement [17]. Training new employees is crucial to acquaint them with countering phishing practices and maintain a consistent level of awareness throughout the organization.

6 Limitations and future work

Despite their advantages, focus groups have a few limitations which we were careful to mitigate through purposeful moderation. The discussion might veer into narratives outside the scope of research. Also, dominant speakers might hijack the discussion while some participants might remain silent and not willing to confront others. This requires researchers' facilitation to steer back to the planned agenda and engage participants with contributing. Furthermore, much of the collected data is expressed informally, necessitating careful interpretation by researchers. Thus, we involved multiple researchers in the data analysis process. Participants' viewpoints might be influenced by the others' arguments during group interaction. Thus, we recorded individual opinions prior to the group discussion on reporting to obtain individual viewpoints.

Although we utilized diverse strategies to recruit employees from the organization, we might have attracted people who are particularly interested in the topic. We hypothesized that an important power imbalance exists between the IT security team and other staff regarding the topic of the study. We did not have IT security officers as participants. We acknowledge that focus groups were composed of participants with multiple roles, potentially creating a perceived power imbalance that inhibited participation. The investigated university has no strict rules regarding phishing awareness campaigns, reporting, and the use of personal devices for work. Thus, while our findings offer valuable insights, critical interpretation is warranted when extrapolating results to different organizational contexts. Future studies should use quantitative methodologies to test the hypotheses drawn from our results.

We found that contextual ("situated") factors, such as task overload, time pressure and stress, influence employees' response to phishing emails (in line with [20]). Contextual factors are not represented in the original EVT framework, although the authors later highlighted that the processes underlying the EVT model are influenced by the immediate situation in which a decision occurs [25]. Recent early-stage work suggests using knowledge about momentary user states to better tailor security interventions [5], for example proposing security interventions or training in opportune moments. We suggest future studies investigate how to integrate contextual factors into EVT when applying it to study information security behaviors.

7 Conclusion

Employees are the last line of organizational defense against phishing attacks [83]. It is important to train and engage employees and encourage reporting of phishing attacks to enable organizations to respond promptly. This engagement can be achieved by enhancing the perceived value of the task, reducing its relative costs, and making *phishing awareness campaigns* more user-centric and relevant to employees.

We find that Expectancy-Value Theory is a valuable theoretical framework for studying user security behavior in an organizational context. EVT helps explain how organizational culture, social roles, and the influence of colleagues and supervisors foster proactive responses to phishing attacks.

Our study reveals a spectrum of factors that influence employees' intentions to *report phishing emails*. Some factors not previously discussed in phishing studies include those associated with social roles (safeguarding the workplace, sense of belonging, and collaboration with IT) and intrinsic factors (satisfaction, enjoyment, and empowerment). Among the factors discouraging employees, the absence of feedback and perceived low utility value are particularly detrimental. This lack not only affects the perceived value of reporting but also undermines employees' confidence in the effectiveness of countermeasures. Given that users devote considerable time and effort in addition to their role to engage in security tasks, it seems justifiable to provide them with more feedback about how their actions fortify the organization's defenses against phishing attacks. A month after our focus group session, we received an email from P18—a highly motivated employee who indicated that they always report suspicious emails. They allowed us to cite:

> *I have now finally stopped reporting phishing emails. Yesterday, I received two that were exactly like the ones I've been getting dozens of times over the past years. It feels a bit like an insult to be asked to report phishing emails when this information is so evidently not utilized. I expressed this sentiment in my final report, but of course, it was ignored.*

We see this loss of engagement with phishing reporting as an understandable but regrettable behavioral response. Envisioning such sentiments and the resulting behavior at scale, with possibly large numbers of employees ending up disappointed and disengaging from phishing interventions, we can only speculate regarding the negative effects on the organizational security of an organization. We hope that this paper can help avoid such frustrating experiences for employees in the future by providing a better understanding of the motivating and discouraging factors for phishing interventions through the lens of EVT.

Acknowledgments

References

[1] ADAMS, A., AND SASSE, M. A. Users are not the enemy. *Communications of the ACM 42*, 12 (1999), 40–46.

[2] ALEROUD, A., AND ZHOU, L. Phishing environments, techniques, and countermeasures: A survey. *Computers & Security 68* (2017), 160–196.

[3] ALHELALY, Y., DHILLON, G., AND OLIVEIRA, T. When expectation fails and motivation prevails: the mediating role of awareness in bridging the expectancy-capability gap in mobile identity protection. *Computers & Security 134* (2023), 103470.

[4] ALKHALIL, Z., HEWAGE, C., NAWAF, L., AND KHAN, I. Phishing attacks: A recent comprehensive study and a new anatomy. *Frontiers in Computer Science 3* (2021), 563060.

[5] ALT, F., HASSIB, M., AND DISTLER, V. Human-centered behavioral and physiological security. In *Proceedings of the 2023 New Security Paradigms Workshop* (New York, NY, USA, 2023), NSPW '23, Association for Computing Machinery, p. 48–61.

[6] BADA, M., SASSE, A. M., AND NURSE, J. R. Cyber security awareness campaigns: Why do they fail to change behaviour? *arXiv preprint arXiv:1901.02672* (2019).

[7] BAGUI, S., NANDI, D., BAGUI, S., AND WHITE, R. J. Machine learning and deep learning for phishing email classification using one-hot encoding. *Journal of Computer Science 17* (2021), 610–623.

[8] BRUNKEN, L., BUCKMANN, A., HIELSCHER, J., AND SASSE, M. A. "To Do This Properly, You Need More Resources": The Hidden Costs of Introducing Simulated Phishing Campaigns. In *32nd USENIX Security Symposium (USENIX Security 23)* (2023), pp. 4105–4122.

[9] BULLEE, J.-W., AND JUNGER, M. How effective are social engineering interventions? a meta-analysis. *Information & Computer Security 28*, 5 (2020), 801–830.

[10] BURNS, A., JOHNSON, M. E., AND CAPUTO, D. D. Spear phishing in a barrel: Insights from a targeted phishing campaign. *Journal of Organizational Computing and Electronic Commerce 29*, 1 (2019), 24–39.

[11] CHANG, C. L.-H., CHEN, V., KLEIN, G., AND JIANG, J. J. Information system personnel career anchor changes leading to career changes. *European Journal of Information Systems 20*, 1 (2011), 103–117.

[12] CHEN, X., SACRÉ, M., LENZINI, G., GREIFF, S., DISTLER, V., AND SERGEEVA, A. The effects of group discussion and role-playing training on self-efficacy, support-seeking, and reporting phishing emails: Evidence from a mixed-design experiment. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2024), pp. 1–21.

[13] CLARKE, V., AND BRAUN, V. Thematic analysis. *The journal of positive psychology 12*, 3 (2017), 297–298.

[14] COOK, D. A., AND ARTINO JR, A. R. Motivation to learn: an overview of contemporary theories. *Medical education 50*, 10 (2016), 997–1014.

[15] CRAM, W. A., D'ARCY, J., AND PROUDFOOT, J. G. Seeing the forest and the trees: a meta-analysis of the antecedents to information security policy compliance. *MIS quarterly 43*, 2 (2019), 525–554.

[16] CRANOR, L. F. A framework for reasoning about the human in the loop.

[17] DA VEIGA, A., ASTAKHOVA, L. V., BOTHA, A., AND HERSELMAN, M. Defining organisational information security culture—perspectives from academia and industry. *Computers & Security 92* (2020), 101713.

[18] DAHABIYEH, L. Factors affecting organizational adoption and acceptance of computer-based security awareness training tools. *Information & Computer Security 29*, 5 (2021), 836–849.

[19] DANG-PHAM, D., KAUTZ, K., HOANG, A.-P., AND PITTAY-ACHAWAN, S. Identifying information security opinion leaders in organizations: Insights from the theory of social power bases and social network analysis. *Computers & Security 112* (2022), 102505.

[20] DISTLER, V. The influence of context on response to spear-phishing attacks: an in-situ deception study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023), pp. 1–18.

[21] DISTLER, V., FASSL, M., HABIB, H., KROMBHOLZ, K., LENZINI, G., LALLEMAND, C., KOENIG, V., AND CRANOR, L. F. Empirical research methods in usable privacy and security. In *Human Factors in Privacy Research*. Springer International Publishing Cham, 2023, pp. 29–53.

[22] DODGE JR, R. C., CARVER, C., AND FERGUSON, A. J. Phishing for user security awareness. *computers & security 26*, 1 (2007), 73–80.

[23] ECCLES, J. Expectancies, values and academic behaviors. *Achievement and achievement motives* (1983).

[24] ECCLES, J. S., AND WIGFIELD, A. Motivational beliefs, values, and goals. *Annual review of psychology 53*, 1 (2002), 109–132.

[25] ECCLES, J. S., AND WIGFIELD, A. From expectancy-value theory to situated expectancy-value theory: A developmental, social cognitive, and sociocultural perspective on motivation. *Contemporary educational psychology 61* (2020), 101859.

[26] ECCLES, J. S., AND WIGFIELD, A. The development, testing, and refinement of eccles, wigfield, and colleagues' situated expectancy-value model of achievement performance and choice. *Educational Psychology Review 36*, 2 (2024), 1–29.

[27] FBI. Internet crime report 2022. https://www.ic3.gov/Media/PDF/AnnualReport/2022_IC3Report.pdf, 2023. Accessed: 10-02-2024.

[28] FORMOSA, P., WILSON, M., AND RICHARDS, D. A principlist framework for cybersecurity ethics. *Computers & Security 109* (2021), 102382.

[29] FRANK, M., AND KOHN, V. Understanding extra-role security behaviors: An integration of self-determination theory and construal level theory. *Computers & Security 132* (2023), 103386.

[30] FRANZ, A. Why do employees report cyber threats? comparing utilitarian and hedonic motivations to use incident reporting tools. In *ICIS 2022 Proceedings* (2022), pp. 1–13.

[31] FRANZ, A., ZIMMERMANN, V., ALBRECHT, G., HARTWIG, K., REUTER, C., BENLIAN, A., AND VOGT, J. Sok: Still plenty of phish in the sea—a taxonomy of user-oriented phishing interventions and avenues for future research. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)* (2021), pp. 339–358.

[32] FUJS, D., MIHELIČ, A., AND VRHOVEC, S. L. The power of interpretation: Qualitative methods in cybersecurity research. In *Proceedings of the 14th International Conference on Availability, Reliability and Security* (2019), pp. 1–10.

[33] GHAZI-TEHRANI, A. K., AND PONTELL, H. N. Phishing evolves: Analyzing the enduring cybercrime. *Victims & Offenders 16*, 3 (2021), 316–342.

[34] GIOIA, D. A., CORLEY, K. G., AND HAMILTON, A. L. Seeking qualitative rigor in inductive research: Notes on the gioia methodology. *Organizational research methods 16*, 1 (2013), 15–31.

[35] GUEST, D. E., AND CONWAY, N. Communicating the psychological contract: an employer perspective. *Human resource management journal 12*, 2 (2002), 22–38.

[36] HAAG, S., SIPONEN, M., AND LIU, F. Protection motivation theory in information systems security research: A review of the past and a road map for the future. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems 52*, 2 (2021), 25–67.

[37] HAN, J., KIM, Y. J., AND KIM, H. An integrative model of information security policy compliance with psychological contract: Examining a bilateral perspective. *Computers & Security 66* (2017), 52–65.

[38] HATTIE, J., HODIS, F. A., AND KANG, S. H. Theories of motivation: Integration and ways forward. *Contemporary Educational Psychology 61* (2020), 101865.

[39] HIELSCHER, J., MENGES, U., PARKIN, S., KLUGE, A., AND SASSE, M. A. "Employees Who Don't Accept the Time Security Takes Are Not Aware Enough": The CISO View of Human-Centred Security. In *32nd USENIX Security Symposium (USENIX Security 23)* (2023), pp. 2311–2328.

[40] HOBBS, A. The colonial pipeline hack: Exposing vulnerabilities in us cybersecurity. In *SAGE Business Cases*. SAGE Publications: SAGE Business Cases Originals, 2021.

[41] HOSSEINI, M., ABDOLVAND, N., AND HARANDI, S. R. Two-dimensional analysis of customer behavior in traditional and electronic banking. *Digital Business 2*, 2 (2022), 100030.

[42] HU, S., HSU, C., AND ZHOU, Z. Security education, training, and awareness programs: Literature review. *Journal of Computer Information Systems 62*, 4 (2022), 752–764.

[43] HULL, D. M., SCHUETZ, S. W., AND LOWRY, P. B. Tell me a story: The effects that narratives exert on meaningful-engagement outcomes in antiphishing training. *Computers & Security 129* (2023), 103252.

[44] JENKINS, J. L., DURCIKOVA, A., ROSS, G., AND NUNAMAKER JR, J. F. Encouraging users to behave securely: Examining the influence of technical, managerial, and educational controls on users' secure behavior. In *ICIS 2010 Proceedings. 150.* (2010).

[45] KENDZIERSKI, D., AND WHITAKER, D. J. The role of self-schema in linking intentions with behavior. *Personality and Social Psychology Bulletin 23*, 2 (1997), 139–147.

[46] KERSTEN, L., BURDA, P., ALLODI, L., AND ZANNONE, N. Investigating the effect of phishing believability on phishing reporting. In *2022 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)* (2022), IEEE, pp. 117–128.

[47] KIRLAPPOS, I., PARKIN, S., AND SASSE, M. A. Learning from "shadow security": Why understanding non-compliance provides the basis for effective security. In *Proceedings of Workshop on Usable Security 2014* (2014).

[48] KRANZ, J., AND HAEUSSINGER, F. Why deterrence is not enough: The role of endogenous motivations on employees' information security behavior. In *International Conference on Information Systems* (2014), IEEE, pp. 1–14.

[49] KUMARAGURU, P., CRANSHAW, J., ACQUISTI, A., CRANOR, L., HONG, J., BLAIR, M. A., AND PHAM, T. School of phish: a real-world evaluation of anti-phishing training. In *Proceedings of the 5th Symposium on Usable Privacy and Security* (2009), pp. 1–12.

[50] KUMARAGURU, P., RHEE, Y., SHENG, S., HASAN, S., ACQUISTI, A., CRANOR, L. F., AND HONG, J. Getting users to pay attention to anti-phishing education: Evaluation of retention and transfer. In *Proceedings of the Anti-Phishing Working Groups 2nd Annual ECrime Researchers Summit* (New York, NY, USA, 2007), eCrime '07, Association for Computing Machinery, p. 70–81.

[51] KWAK, Y., LEE, S., DAMIANO, A., AND VISHWANATH, A. Why do users not report spear phishing emails? *Telematics and Informatics 48* (2020), 101343.

[52] LAIN, D., KOSTIAINEN, K., AND ČAPKUN, S. Phishing in organizations: Findings from a large-scale and long-term study. In *2022 IEEE Symposium on Security and Privacy (SP)* (2022), IEEE, pp. 842–859.

[53] LEHRE, F. . Hochschulen im visier von cyberkriminellen. https://www.forschung-und-lehre.de/management/hochschulen-im-visier-von-cyberkriminellen-5541, 2023. Accessed: 10-02-2024.

[54] LI, Y., STAFFORD, T. F., FULLER, B., AND ELLIS, S. Beyond compliance: Empowering employees' extra-role security behaviors in dynamic environments. In *AMCIS* (2017).

[55] MANSFIELD-DEVINE, S. Raising awareness: People are your last line of defence. *Computer Fraud & Security 2017*, 11 (2017), 10–14.

[56] MARIN, I. A., BURDA, P., ZANNONE, N., AND ALLODI, L. The influence of human factors on the intention to report phishing emails. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023), pp. 1–18.

[57] MENARD, P., BOTT, G. J., AND CROSSLER, R. E. User motivations in protecting information security: Protection motivation theory versus self-determination theory. *Journal of Management Information Systems 34*, 4 (2017), 1203–1230.

[58] POSEY, C., ROBERTS, T., LOWRY, P. B., COURTNEY, J., AND BENNETT, B. Motivating the insider to protect organizational information assets: Evidence from protection motivation theory and rival explanations. In *The Dewald Roode workshop in information systems security* (2011), pp. 22–23.

[59] RAHIMI, S., AND KHATOONI, M. Saturation in qualitative research: An evolutionary concept analysis. *International Journal of Nursing Studies Advances 6* (2024), 100174.

[60] REEVES, A., CALIC, D., AND DELFABBRO, P. "generic and unusable" 1: Understanding employee perceptions of cybersecurity training and measuring advice fatigue. *Computers & Security 128* (2023), 103137.

[61] RENAUD, K., SEARLE, R., AND DUPUIS, M. Shame in cyber security: effective behavior modification tool or counterproductive foil? In *New Security Paradigms Workshop* (2021), pp. 70–87.

[62] RIZZONI, F., MAGALINI, S., CASAROLI, A., MARI, P., DIXON, M., AND COVENTRY, L. Phishing simulation exercise in a large hospital: A case study. *Digital Health 8* (2022), 20552076221081716.

[63] ROGERS, R. W. A protection motivation theory of fear appeals and attitude change1. *The journal of psychology 91*, 1 (1975), 93–114.

[64] SASSE, M. A., HIELSCHER, J., FRIEDAUER, J., AND BUCKMANN, A. Rebooting it security awareness–how organisations can encourage and sustain secure behaviours. In *European Symposium on Research in Computer Security* (2022), Springer, pp. 248–265.

[65] SILIC, M., AND LOWRY, P. B. Using design-science based gamification to improve organizational security training and compliance. *Journal of Management Information Systems 37*, 1 (2020), 129–161.

[66] SOLOMON, G., AND BROWN, I. The influence of organisational culture and information security culture on employee compliance behaviour. *Journal of Enterprise Information Management 34*, 4 (2021), 1203–1228.

[67] SON, J.-Y. Out of fear or desire? toward a better understanding of employees' motivation to follow is security policies. *Information & Management 48*, 7 (2011), 296–302.

[68] THOMAS, A., AND GUPTA, V. The role of motivation theories in knowledge sharing: an integrative theoretical reviews and future research agenda. *Kybernetes 51*, 1 (2022), 116–140.

[69] UCHICAGO. Latest phishing scams. https://security.uchicago.edu/phishing/latest/, 2024. Accessed: 10-02-2024.

[70] VANCE, A., SIPONEN, M., AND PAHNILA, S. Motivating is security compliance: Insights from habit and protection motivation theory. *Information & Management 49*, 3-4 (2012), 190–198.

[71] VERBISOFTWARE. Maxqda. https://www.maxqda.com/, 2022. Accessed: 10-02-2024.

[72] VOLKAMER, M., RENAUD, K., REINHEIMER, B., RACK, P., GHIGLIERI, M., MAYER, P., KUNZ, A., AND GERBER, N. Developing and evaluating a five minute phishing awareness video. In *Trust, Privacy and Security in Digital Business: 15th International Conference, TrustBus 2018, Regensburg, Germany, September 5–6, 2018, Proceedings 15* (2018), Springer, pp. 119–134.

[73] WARKENTIN, M., JOHNSTON, A. C., AND SHROPSHIRE, J. The influence of the informal social learning environment on information privacy policy compliance efficacy and intention. *European Journal of Information Systems 20*, 3 (2011), 267–284.

[74] WEAVER, B. W., BRALY, A. M., AND LANE, D. M. Training users to identify phishing emails. *Journal of Educational Computing Research 59*, 6 (2021), 1169–1183.

[75] WEN, Z. A., LIN, Z., CHEN, R., AND ANDERSEN, E. What. hack: engaging anti-phishing training through a role-playing phishing simulation game. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), pp. 1–12.

[76] WIGFIELD, A., AND ECCLES, J. S. Expectancy–value theory of achievement motivation. *Contemporary educational psychology 25*, 1 (2000), 68–81.

[77] WIGFIELD, A., TONKS, S., AND KLAUDA, S. L. Expectancy-value theory. *Handbook of motivation at school 2* (2009), 55–74.

[78] WILKINSON, S. Focus group methodology: a review. *International journal of social research methodology 1*, 3 (1998), 181–203.

[79] WILLIAMS, E. J., HINDS, J., AND JOINSON, A. N. Exploring susceptibility to phishing in the workplace. *International Journal of Human-Computer Studies 120* (2018), 1–13.

[80] WILLIAMS, P. Organisational culture: definitions, distinctions and functions. *Handbook of research methods for organisational culture* (2022), 5–22.

[81] YEOH, W., HUANG, H., LEE, W.-S., AL JAFARI, F., AND MANSSON, R. Simulated phishing attack and embedded training campaign. *Journal of Computer Information Systems 62*, 4 (2022), 802–821.

[82] ZHENG, S. Y., AND BECKER, I. Phishing to improve detection. In *Proceedings of the 2023 European Symposium on Usable Security* (2023), pp. 334–343.

[83] ZIMMERMANN, V., AND RENAUD, K. Moving from a 'human-as-problem" to a 'human-as-solution" cybersecurity mindset. *International Journal of Human-Computer Studies 131* (2019), 169–187.

A   The core constructs of Expectancy-Value Theory

The core constructs of Expectancy-Value Theory as described in Eccles and Wigfield's work [25] are as follows:

- **Expectation of Success**: Individuals' beliefs regarding their potential effectiveness in executing tasks or resolving challenges [25].
- **Achievement-Related Choices and Performance**: The outcomes that individuals target when they choose to engage with an activity or perform a task, informed by their interpretation of expectation of success and perceived value of the specific task [25].
- **Subjective Task Value**: Individuals' assessment of a task's significance, utility, emotional resonance, and perceived cost [25].
- **Goal**: Cognitive representation of a future outcome that an individual is striving to achieve [24].
- **Self-schemata**: Cognitive generalizations about oneself, derived from past experiences and focused on self regarded importance [45].
- **Affective Reactions and Memories**: Individuals' emotional responses to specific tasks or scenarios, alongside the emotive memories derived from past experiences [76].

- **Perception of**: Individuals' interpretation and understanding of their previous experiences and socialization influences [76].
- **Interpretation of Experience**: The personal lens through which individuals perceive prior achievement-related events, influenced by a confluence of cultural, social, external feedback, and intrinsic cognitive and emotional factors [76].
- **Cultural Milieu**: A system of social roles, each with its associated responsibilities and obligations [77], this construct has been extended in our study to encompass "organizational culture."
- **Socializer**: Originally pertaining to parents, educators, and extended social circles in EVT [76], this construct has been adapted in our context to also include "colleagues and supervisors."
- **Person Characteristics**: The array of individual variances, encapsulating aspects such as abilities, personality dimensions, gender, age, and cultural origins [76].
- **Previous Achievement-Related Experiences**: Individuals' past experiences in activities or tasks that had a measurable outcome [25].

B   The templates and focus group protocol

Introduction: Thank you for participating in this focus group discussion. This study is one part of the "anonymized" project, funded by "anonymized". This focus group aims to learn about employees' participation in and opinions on phishing awareness campaigns and reporting suspicious emails.

During the discussion, we will record audio and video and collect the paper materials. The collected data will only be used for this study. You have the right to access, rectify, and erase your data. Your participation in the project is voluntary; you can withdraw at any point without giving reasons. You may skip any task you do not wish to participate in for any reason, at any time, without explanation.

There are no right or wrong answers to the questions we prepared; also, we will not ask you questions about your passwords or whether you have encountered phishing attacks in the past. All your answers will be kept strictly confidential and will be anonymized, encrypted and only reviewed by the researchers of this project. Any data shown externally, for example in publications or presentations, will also be anonymized. Your data will be stored and processed only for the purpose of the study stated above for a period of 63 months on internal, on-premises servers.

The focus group will take approximately 90 minutes. Each participant will be compensated with a 40-euro voucher for participation. Do you have any questions so far? If you agree with the terms, please sign the consent form, and then we can start the recording and begin the focus group discussion. The focus group includes four main parts: warm-up activity,

Figure 2: Template 1, what motivates and discourages you in a leisure activity.

Figure 3: Template 2, what motivates and discourages you in reporting.

discussion, brainstorming and debriefing. Let's first have the warm-up activity.

Part 1: Warm-up activity (10 minutes):

Icebreaker: Now, you have 2 minutes to observe the items presented in the lab, try to spot one item that can be used to describe you today. We will share our thoughts after 2 minutes.

Explore motivational and discouraging factors for a leisure activity: Great, now we know each other. Let's move on to explore factors that motivate and discourage you from engaging in a leisure activity. You have 5 minutes to answer the questions on Template 1 (see Figure 2). After you finish, we will collect the paper.

Part 2: Group discussion (60 minutes):

Now, let's move on to the discussion session. Phishing attack is a type of social engineering attack where attackers send spoofed or deceptive messages to trick a person into revealing sensitive information to the attacker or to deploy malicious software on the recipient's devices. Currently at our organization, we have several practices to raise employees' awareness of phishing attacks. First, the IT department sends simulated phishing emails to employees to raise awareness of potential phishing attacks. Second, our university has purchased online security courses from a service provider; you can access the learning platform via this link: "Anonymized". Third, the IT department distributes posters and sends emails to inform employees of online security courses. Some of you might have received these emails or saw the posters at the entrance to the administrative buildings.

Discuss phishing awareness campaigns:

1. What do you think of these three phishing campaigns offered by the IT team?
2. What are the benefits of participating in phishing campaigns?
3. What are the costs of participating in phishing campaigns?
4. Assuming that you know how to take the online security courses, what would discourage you from taking these courses?

5. Have you set any goals for yourself in terms of defending yourself from phishing attacks?

6. How confident are you in protecting yourself from phishing attacks?

Thank you for sharing these opinions with us. In our university, the IT department recommends that employees report phishing emails to report-a-phish@"anonymized"; the Outlook client now also has a report phishing emails button, so you can report with one click.

Now, you have five minutes to fill-in Template 2 (see Figure 3), "what motivates and discourages you from reporting suspicious emails"... Thank you and let's move on to **discuss reporting suspicious emails**:

1. The IT department suggests that we report phishing emails, what do you think of this suggestion?

2. How confident are you about identifying and reporting suspicious emails?

3. As a member of the organization, how do you see your role in reporting suspicious emails?

4. What are the benefits of reporting suspicious emails?

5. What are the costs of reporting suspicious emails?

6. What would discourage you from reporting phishing emails?

Part 3: Brainstorming (15 minutes): Assume that you are our university's new chief information security officer (CISO), and you learned that there are increasing phishing emails targeted at our university. What would you do to motivate employees to engage with these counter-phishing practices?

Part 4: Debriefing (5 minutes): Introduce the IT department recommendations of participating in phishing awareness campaigns and reporting suspicious emails.

C Coding scheme and exemplar quotes

C.1 Factors associated with phishing awareness campaigns

C.1.1 Motivating factors

Gaining phishing knowledge: Participants learned about the techniques and tricks of phishing attacks.

If you were participating in this awareness campaign, maybe get to know some new tricks and what is going on. Maybe there are new types of phishing. (P3)

Acquiring skills: Employees acquired skills in identifying whether the emails, links and website URLs are legitimate or not.

(Phishing campaigns)... train people to recognize what is phishing and prevent them from actually falling into one when it happens. (P22)

Enhancing phishing awareness: The phishing campaign raised employees' awareness of phishing attempts and made them more vigilant against potential attacks.

The good thing is if we make mistakes, they don't cost anything because they're internal mistakes. But they raise our awareness. (P26)

Cyber safety: Participants felt better prepared to protect themselves, their emails, and their workplace from phishing attacks.

It not only benefits you because you will protect your data and your e-mail accounts and so on; will also help the university as an institution to be better protected. (P9)

Personal development: Participants believed that the knowledge gained could benefit their daily life.

It's not only about fear of being attacked, you need to understand what's inside these technology tools... Everything related to cybersecurity is very fundamental now and, in the future, would become even more fundamental, like reading. (P29)

C.1.2 Discouraging factors

Perceived low value: Participants assumed that online phishing courses only provide very basic knowledge or use too complex terms for them to understand.

Don't give me a half hour course for two minutes' value. (P13)

Lack of interest: Negative impressions of the courses, such as "not interesting" and "too easy".

They look like really boring corporate mandated trainings and also the title "Anonymized", look at that and I'd be like oh no... (P17)

Secondary task: Participants mentioned that the phishing campaign was not relevant to their area of expertise or job position.

My role is more task oriented. So, I have to finish my tasks by the end of the day. If I take a course that's one hour long, that means I leave one hour later. (P24)

Lack of incentive: Participants considered lack of incentives, such as course credits, compensation, or praise from the team leader, as discouraging engagement with the awareness campaign.

What is my incentive to do an optional course here? (P24)

Time: Participants mentioned time as a constraint that discourages them from engaging in phishing campaigns.

Sometimes when you are busy, it's very hard to find an hour or so in a day to do them, and so it's quite a big constraint on that. I would say it's mainly time. (P9)

Interrupting workflow: Participating in awareness campaigns required people to switch away from the task at hand to phishing-related content.

The cost is the time spent, but also entering into the actual narrative and that type of discourse. Because you're doing something else and then you're switching to this. And you're like, OK, it's a completely different world, so it takes you away from your attention span. (P25)

Optimism bias: Participants mentioned that they believed they were less likely to fall for phishing than others.

I always had this thinking like, it won't happen to me because this is so stupid. (P14)

Overconfidence: Participants stated that they are very confident in their knowledge of the topic.

I should spend my time doing something else so it's like a prerequisite of this course like ... like 70 to 80% of course material they have already known. (P21)

Procrastination: Participants shared that procrastination resulted in delaying or "forgetting to" take the courses.

If there's no deadline, if there's no shock, I'll do it tomorrow, tomorrow, tomorrow. (P32)

Negative inference: Participants would become more worried about all the potential threats they might receive if they participated in awareness campaigns.

More negative inference ... we become a bit more scared about all these potential threats that we might receive. A little bit of stress in a sense that we need to be careful. (P30)

Fear of failing training: Previous bad experiences with awareness campaigns might evoke fear of failing the training.

The fear or the worry that if I failed the course, it would be tracked. Because I experienced that in the previous job. If you didn't get a certain grade, then you would be forced to retake it and retake it. (P8)

C.2 Factors associated with reporting

C.2.1 Motivating factors

Collaborating with the IT team: Participants considered reporting as a collaboration with the IT team. The IT team assists the employees in verifying the legitimacy of the emails, and employees assist the IT team in detecting the phishing attempts in real-time.

I think this is essential that we can report phishing to IT; and based on that they can have some statistics and see how the attacks are evolving. (P5)

Safeguarding the workplace: Participants regarded reporting as a measure to protect their workplace and colleagues.

Safeguard yourself, your institution, because I'm aware of phishing attacks that cause huge damages in the banking and insurance sector, in research departments overseas, and it's reputational damage that I would not like to be associated with. So protection for the whole institution and for me ultimately. (P13)

Expectation of mitigation: Participants expected that the organization would improve its spam filters and mitigate the attack promptly with the reported emails.

The main benefit of reporting is that the IT team could create more filters for phishing emails if they have more data (from reporting), making us safer (P27)

Recognition: Participants regarded the "congratulations" email they received from the IT team as a kind of recognition

and extrinsic reward for their reporting.

And personally, it's always nice to have, like the congratulations, it's a nice accomplishment and you have the impression that you'd be helping the university community, so it's kind of rewarding. (P9)

Fear of consequences: Worries and fears related to not reporting prompt participants to report phishing attempts.

There're serious consequences if a phishing goes through, from a company perspective or on a personal level. (P13)

Sense of belonging: Participants expressed being part of the community prompts them to engage in reporting phishing.

We need to participate. We're all, we're all active users and it's not just IT who has to deal with it. (P32)

We are actors within the community. So, we are together. (P34)

Responsibility: Participants regarded reporting phishing as part of their job and shared the responsibility of reporting.

I see my role as a little more than this reporting, but also trying to reduce all the risk ... we have a duty. And you owe it to your colleagues as well as yourself. (P11)

Peer influence: Participants reported phishing emails because of the influence of their colleagues.

I used to ignore these emails, but then like one of my colleagues told me, it's better to report. So then I started doing it, yeah, but even I don't do it like every time, but most of the time I try to report them. (P21)

Easy to report: Participants mentioned that the positive user experience with the reporting process motivates them to report.

The reporting button is really easy, even if you're in doubt, you tend to click the button. (P13)

Protecting oneself: Participants considered reporting to benefit them in protecting personal accounts, avoiding financial losses, and safeguarding data.

If I never report anything, I can't expect it to just magically get better, so that's why I see a benefit for myself. (P26)

Phishing experience: Participants mentioned their experiences with phishing incidents as a driver for reporting.

I had this scam attack, and I felt bad about myself. I felt bad about trusting the others, so I wouldn't like someone, other people to feel the same way I felt once. (P4)

Empowerment: Participants considered reporting as an initiative against phishing attempts, giving them a sense of control and empowerment.

I had the initiative to defend against the phishing attack. And knowing that I can stop spreading this attack for other people and for my future self. That really helped me, like empowering. (P16)

Satisfaction: Participants expressed their sense of accomplishment/satisfaction for reporting suspicious emails.

I can relate to the sense of satisfaction. Once you've reported it, you feel like you played your role. You did a good job. (P11)

Enjoyment: Participants considered the reporting as a playful game or "nice welcome distraction" from work.

When you click to report phishing attempts, then you receive 'congratulations'. I'm happy and it's like a game. (P28)

Personal Value: Participants reported phishing attempts because it is the right thing to do or the suggestion is good.

It's a very good action to ask us to report suspicious emails. (P6)

Altruism: Participants wanted to help others and vulnerable groups, reducing their chances of being phished.

I want to help others avoid being deceived by phishing. (P15)

Pride: Participants mentioned pride stemming from their ability to consistently identify and avoid being phished.

I don't want to break my streak of always reporting the phishing attacks. I've not clicked on one socially engineered phishing e-mail, I'm quite proud of that. (P8)

C.2.2 Discouraging factors

Perceived low threat: If the participants regarded the incoming phishing emails as too obvious/low threat, they chose not to report.

If I consider the content of phishing emails so apparent, so explicit that everyone can find out that it's phishing, then I don't try to report it. (P16)

Negative outcomes: Assumed negative outcomes from reporting the email discouraged participants.

I feel like there's negative benefits for me reporting them because they don't seem to do anything with it and I just get more emails. So I would get the same amount of spam if I didn't report it. (P17)

Report too much: Participants expressed the concern that they reported too many suspicious emails and burdened the IT team.

It's already the second one I sent this week, so I said, what shall I do? (P28)

Worries of being judged: Participants expressed reservations about reporting suspicious emails due to worries of being judged by the IT team.

If I report Netflix or something as phishing, then they would think 'stupid woman'... This feeling unnerved me and discouraged me from reporting. (P34)

Privacy concerns: Participants expressed they were hesitant about reporting when they felt that it might divulge private information or create a false impression about their personal life.

I worry what they (the IT team) will think of me. So, I try to avoid informing them, because what are they doing with this information? (P28)

Switching between interfaces: Participants mentioned that even they intended to report suspicious emails, they tended to delete or ignore them when checking email on their smartphone.

I use the web client sometimes. I don't know if there is a report phishing on there, and I also don't know if it's on like the iPhone app. (P24)

Unclear procedures: Participants shared that unclear reporting procedures discouraged them from reporting suspicious emails.

I think you should report the suspicious emails, but it needs to be made clearer what suspicious e-mail is and how to properly report it. (P8)

Requiring too much effort: Participants who use Linux and Mac OS expressed that the reporting procedure requires too much effort.

It's too much effort for me, like not much effort, but it's not very easy. (P27)

Lack of feedback: Without follow-up or feedback on their reporting action, participants felt discouraged from reporting.

We don't know what the effectiveness of report-a-phish is. We don't know the numbers, so it would be really good to have a kind of feedback status. What has been done last year? What was the success rate? (P31)

Lack of communication: Participants felt discouraged due to not knowing whether their colleagues reported or not and the organization's status quo for reporting.

I report phishing emails regularly and religiously, but I'm thinking is everyone else doing the same as me, putting in the same effort as I am on reporting? It takes maybe 30 seconds of your time, butt I'm still very careful about it. (P25)

Low response efficacy: When they perceived no impactful results of their reporting, participants felt discouraged and even stop reporting.

If we feel it works, maybe we continue to report, but if it does not work so well, we will not report phishing again. (P1)

Habitual behavior: Participants shared that they often postponed or forgot to report because they reverted back to old habits of simply deleting emails.

Just going back to your old habits because this report phishing button for me is new. And in my other like personal e-mail, Gmail, what I do is delete. So, I might result in just deleting and then other times I might remember. (P11)

Laziness: Participants mention "laziness" as a self-reported reason for not reporting suspicious emails.

I'm able to report them, but sometimes I'm too lazy to report it. (P17)

Low self-efficacy: If they had too high doubts and were not confident about whether it was a phishing attempt or not, participants would not report.

For reporting, I'm not sure because sometimes I am not sure it indeed is a phish or not, so then sometimes, I just prefer to delete it and not to report. (P5)

Simulated or real attack: When simulated phishing tests are overused or not accompanied by a clear protocol, they result in reduced reporting intentions.

For me, every phishing email that I received was a simulated one. So, I didn't see the point of reporting that because

I knew that it was from IT. (P27)

Contextual factors: Overload at work, time pressure and stress when they received the email could discourage them from reporting.

Sometimes when I'm in a rush, I just delete. (P31)

D   The demographic table

Table 2: Demographic table of focus groups.

Focus group	Participant	Job title	Field	Work experience (years)[a]
FG01	P1	Doctoral researcher	Computer Science	1
	P2	Lead software developer	IT	21
	P3	Doctoral researcher	Energy	4
	P4	Doctoral researcher	Robotics	12
	P5	Postdoctoral researcher	Security and cryptography	5
FG02	P6	Doctoral researcher	Psychology	7
	P7	Doctoral researcher	Psychology	2
	P8	Administrative assistant	Administration	2
	P9	Doctoral researcher	Political science and human rights	5
FG03	P10	Doctoral researcher	Neuroscience	5
	P11	Doctoral researcher	Social economics	5
	P12	Postdoctoral researcher	Engineering	8
	P13	Postdoctoral researcher	Digital health	23
	P14	Doctoral researcher	Political sciences	5
	P15	Doctoral researcher	Law	3
	P16	Doctoral researcher	Social sciences	1
FG04	P17	Doctoral researcher	Computer Science	5
	P18	Software developer	IT	20
	P19	Doctoral researcher	Computer Science	8
FG05	P20	Administrative assistant	Administration	25
	P21	Doctoral researcher	Supply chain management	2
	P22	Postdoctoral researcher	Security and cryptography	5
	P23	Doctoral researcher	Engineering	3
FG06	P24	Building project manager	Administration	13
	P25	Academic facilitator	Administration	26
	P26	Alumni relations	Administration	34
	P27	Software developer	IT & Admin	23
FG07	P28	Research facilitator	Administration	30
	P29	Data analyst	Administration	7
	P30	Research facilitator	Administration	21
	P31	Project manager	Administration	25
	P32	Research facilitator	Administration	27
	P33	Secretary	Administration	30
	P34	Administrative assistant	Administration	35

[a] We removed gender, age, and months working at the current organization to avoid re-identification. Work experience indicates the participants' total years of work experience, including previous jobs.

Beyond the Office Walls: Understanding Security and Shadow Security Behaviours in a Remote Work Context

Sarah Alromaih[1,2], Ivan Flechais[1], George Chalhoub[1,3]

[1]*University of Oxford, Oxford, UK*
[2]*King Abdulaziz City for Science and Technology, Riyadh, Saudi Arabia*
[3]*University College London, London, UK*
[1]*{sarah.alromaih, ivan.flechais}@cs.ox.ac.uk,* [3]*g.chalhoub@ucl.ac.uk*

Abstract

Organisational security research has primarily focused on user security behaviour within workplace boundaries, examining behaviour that complies with security policies and behaviour that does not. Here, researchers identified shadow security behaviour: where security-conscious users apply their own security practices which are not in compliance with official security policy. Driven by the growth in remote work and the increasing diversity of remote working arrangements, our qualitative research study aims to investigate the nature of security behaviours within remote work settings.

Using Grounded Theory, we interviewed 20 remote workers to explore security related practices within remote work. Our findings describe a model of personal security and how this interacts with an organisational security model in remote settings. We model how remote workers use an appraisal process to relate the personal and organisational security models, driving their security-related behaviours. Our model explains how different levels of alignment between the personal and organisational models can drive compliance, non-compliance, and shadow security behaviour in remote work settings. We discuss the implications of our findings for remote work security and highlight the importance of maintaining informal security communications for remote workers, homogenising security interactions, and adopting user experience design for remote work solutions.

1 Introduction

Organisational security research has primarily focused on user security behaviour within workplace boundaries [42]. User behaviour typically falls into two categories with regard to security policies: those who comply with security policies and those who do not [33]. Within the non-compliant space, researchers have identified shadow security behaviour [38]— where security conscious users come up with their own security practices when they cannot comply with the official security policy.

Along with improvements in collaborative work technologies, the global COVID-19 pandemic pushed individuals outside of organisational perimeters and established remote work as the "new normal". The 2022 workplace trends and insights report [2] revealed that 73% of employees now operate in a hybrid or fully remote setting and nearly half work entirely from home. Interestingly, a third of workers expressed their preference to continue working in a fully remote capacity.

Yet, despite the growing interest in remote working, the existing literature on user security-related behaviour has mostly focused on contexts where remote work is not so prevalent (e.g. [39], [34], [10]). Furthermore, to our knowledge, no user study has been conducted to explore users' security behaviour and shadow practices entirely in the context of remote work. To explore this gap, our overarching research question is: *What are the current security and shadow security practices in remote work?*

To address our research question, we used Grounded Theory [13, 16, 23] to conduct and analyse a qualitative semi-structured interview study with 20 participants engaged in remote work, each employed by a single employer (i.e., an external organisation and not their own business), aiming to explore security related practices within remote work.

Our findings describe three different models which interplay with one another and help describe security practices in remote work. The first consists of a personal security model driven by a variety of external factors, including past experiences, past incidents, qualifications, external advice, and in-

teractions with online services and technologies. The second consists of the current organisational security model which significantly influences the personal model, and consists of security rules and tools disseminated formally through security awareness and training, and informally through interaction with colleagues and the security culture. The third is a model of an appraisal process which individuals use to relate the personal and organisational security models to help them decide which security practices they should follow. This model explains how different levels of alignment between the personal and organisational models can drive compliance, non-compliance, and shadow security behaviour in remote work settings.

In helping to explain security behaviour in remote work, our findings support prior research that notes that shadow security practices can arise from perceptions of inappropriate organisational policies and rules [6, 39]. We discuss the implications of our findings for remote work security, highlighting the challenge of maintaining informal security communications for remote workers to help foster a strong security culture, the need for greater consistency in the experience of security interactions across devices and services, and the wider value of considering the user experience of remote work security in the design of new technology and in the operation of remote work organisations.

The rest of the paper is organised as follows: In Section 2, we give a background overview of related topics. We elaborate on our research methodology in Section 3. We present and discuss our results in Sections 4 and 5, respectively. Finally, we conclude our paper in Section 7.

2 Background

In this section, we will review security compliance and shadow practices within the workplace, followed by an overview of remote work as the context of our research study. Lastly, we will discuss remote work security.

2.1 Security Compliance and Shadow Practices

Security in its simplest form can be described as "*things that should happen, do, and things that shouldn't happen, don't.*" [54]. Therefore, organisations implement various controls and measures to ensure effective security within the workplace. These controls and measures range from technical and non technical solutions to organisational security awareness and training. Among these controls, the information security policy is the most important, since it indicates how workers should behave in order to mitigate security risks [33].

User behaviour typically falls into two categories with regard to security policies: those who comply with security policies and those who do not [29]. Since 1999, Adams and Sasse [3] have noted that for some users it is impossible to meet both security policy requirements and complete their

main work task in a timely manner, leading to further studies to suggest a third category, which is shadow security behaviour [38]— where security conscious users come up with their own security practices when they cannot comply with the official security policy.

Shadow security practices have the same characteristics as shadow Information Technology (IT) phenomenon in that they are both covert and unofficial. Shadow IT refers to any hardware, software, and other solutions employed by users without explicit approval or knowledge from their organisations [30, 31]. There are many terms used in the literature to describe this phenomenon, including shadow IT, shadow systems, rogue IT, workaround systems, grey IT or feral systems [52, 56]. Shadow IT solutions can take the form of a simple Excel spreadsheet [52] or a complex application integrated with the official systems [57]. The proliferation of portable devices, cloud technologies, and subscription-based software or services have transformed traditional IT management and contributed to shadow IT becoming more prevalent [44].

Kirlappos et al. [38] investigated security policy non-compliance by interviewing employees within a large organisation. This study revealed instances of shadow security in which employees create workarounds that try to achieve reasonable security goals as a more suitable alternative to prescribed security policies. The researchers suggested that security experts should take cues from these shadow security practices, given that these practices offer a basis for workable security protocols better aligned with employees' workplace goals [39].

2.2 Remote Work As Context

Remote work, also referred to as telecommuting, telework, flexible work arrangements, distributed work and virtual teams [5], is the ability to work outside of an organisation's physical workplace as part of a flexible working arrangement [1]. With respect to location and time, remote work encompasses various modalities, enabling individuals to work from nearly anywhere—primarily from home, but also from other locations such as communal spaces (e.g., libraries, coffee shops) or co-working environments. This flexibility sometimes includes the option for asynchronous work, allowing employees to select their working hours based on their productivity peaks and personal commitments [27]. Additionally, there is the hybrid working model, where employees blend office days with remote workdays as part of their working arrangement to combine the best of both settings [58].

The concept of remote work, whether from home or while on the move, has been in existence for some time [49]. However, with the improvements in information and communication technologies (ICT), the global COVID-19 pandemic pushed individuals outside of organisational boundaries and established remote work as the "new normal" [53]. As a result, remote work has boomed since the COVID-19 pandemic, in contrast to the steady increase observed between 1980 and

2019 [47]. Furthermore, this trend reflects a growing acceptance among employers in allowing employees to work remotely. According to Hansen et al. research [32], from 2019 to early 2023, the proportion of job postings offering new employees the option to work remotely increased by more than threefold in the U.S. and by a factor of five or more in Australia, Canada, New Zealand, and the UK. This growth has significantly expanded knowledge workers' access to job opportunities and better incomes but also posed cybersecurity challenges, despite security not being a frequent priority in this context [22].

2.3 Security of Remote Work

In 2021, a study by Bispham et al. [9] found a lack of research on cybersecurity in remote work and distance education, despite the extensive use of internet and computing technologies in these domains. The authors conducted exploratory in-depth interviews with cybersecurity experts and remote work support staff. The interviews revealed several security challenges associated with remote work, including an uptick in phishing attacks, a higher number of compromised accounts, and an increase in ransomware attacks.

Researchers and industry experts have proposed various solutions to address cybersecurity risks in this context, such as scaling up the use of virtual private networks (VPNs) and Multi-Factor Authentication (MFA), implementing endpoint protection, providing user education on phishing scams, implementing zero trust model [60], establishing robust policies for mobile device management (MDM), and considering cloud migration strategies to protect organisational assets [20, 43, 51]. Nevertheless, as indicated by the exploratory interviews conducted by Bispham et al. [9], "the best approaches to security are unsettled and evolving".

Godlove [25] provided insights for organisations with remote workers regarding data security attitudes and compliance. A survey of 150 remote workers revealed that personal attitude, social pressure, sense of control, and responsibility moderately explain their willingness to follow security guidelines. Yet, despite the growing interest in remote working, the existing literature on user security-related behaviour has mostly been investigated in contexts where remote work was infrequently practised by only a few employees (e.g. [39], [34], [10]), with no focus on shadow security. Our goal is to address this gap by exploring user security behaviour and shadow practices in the context of remote work.

3 Methodology and Research Question

For this exploratory research study, we adopted a qualitative research design, guided by the constructivist approach to Grounded Theory proposed by Charmaz [13] to address our research question: *What are the current security and shadow security practices in remote work?*

Originally proposed by Glaser and Strauss [24], Grounded Theory has shown to be a well-established methodology for exploring security research [19, 50], and is particularly suited to areas of inquiry that have not been widely researched. Also, it allows examining topics and situations from several perspectives, which can lead to comprehensive and deep explanations. It can uncover underlying perspectives, perceptions, and beliefs that influence behaviours, practices, and incidents by examining both rational and irrational aspects [61]. We designed and conducted semi-structured interviews with 20 participants who were working remotely, either fully remote or in a hybrid mode, and we employed the constructivist approach to Grounded Theory by Charmaz [14] as a data analysis method aiming to construct substantive theory through a structured, flexible, iterative and comparative process of analysing the data [15]. An overview of the research process and applied methods is shown in Figure 1.

3.1 Recruitment and Sampling

To recruit our participants, we adopted purposive sampling to initially identify our target participants. This method was complemented by snowball sampling to further expand the participants group [48]. We advertised the study on online platforms, such as LinkedIn and X (formerly Twitter), aiming to recruit individuals working remotely for a single employer (i.e., an external organisation and not their own business), either fully remote or in a hybrid mode. Also, we expanded our pool of participants by encouraging interested individuals to refer us to suitable contacts from their networks, employing a snowball sampling approach [26].

Interested individuals who met our criteria received a study information sheet and a consent form. Upon signing the consent form, they were requested to complete an online questionnaire regarding their demographic information. The demographic information includes participant age, gender, education, location, organisation business domain, current job role, work settings and level of technical competency in computer security. We defined different levels of technical competence (novice, competent, and expert) using a simplified version of Dreyfus' skill acquisition model that has been widely used to define levels for assessing individual competency [18]. Our demographic information questionnaire can be found in Appendix A.

We interviewed 20 participants: 11 reported working fully remotely, while 9 worked in a hybrid mode. A detailed overview of our sample demographics is presented in Table 1.

3.2 Interview Procedure

We conducted semi-structured interviews with 20 remote workers. We designed and structured our interview guide according to the funnel technique [11], starting with general open-ended questions and gradually moving to specific ones. Using this approach helps build rapport with the participants to clarify and obtain more specific information about their
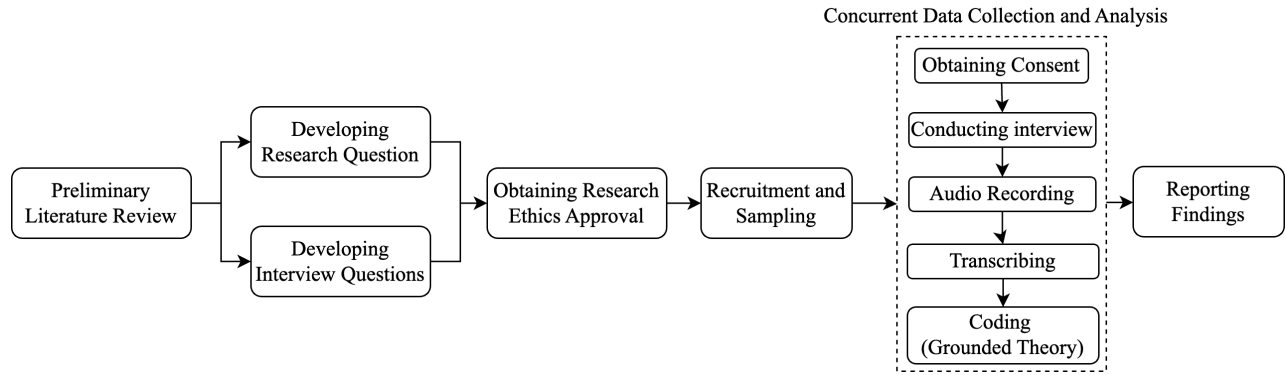
Figure 1: An overview of the research process

P#	Age (M/F)	Degree	Location	Domain	Role (Work Setting)	Competence
P01	25-34 (M)	Postgrad	UK	Tech/IT	Researcher (Remote)	Competent
P02	25-34 (M)	Undergrad	USA	Tech/IT	Products Consultant (Hybrid)	Competent
P03	25-34 (M)	Grad	USA	Tech/IT	Cybersecurity Professional (Hybrid)	Expert
P04	25-34 (F)	Grad	UK	Tech/IT	Product Manager (Remote)	Competent
P05	25-34 (M)	Grad	UK	Consulting	Director (Hybrid)	Competent
P06	25-34 (M)	Grad	UK	Consulting	Associate Software Consultant (Hybrid)	Competent
P07	25-34 (M)	Postgrad	Germany	Tech/IT	Security Awareness Advocate (Remote)	Expert
P08	25-34 (M)	Undergrad	UK	Tech/IT	Full-Stack Developer (Remote)	Expert
P09	35-44 (F)	Grad	UK	Retail	Finance (Remote)	Expert
P10	25-34 (M)	Postgrad	USA	Tech/IT	Editor (Remote)	Novice
P11	35-44 (M)	Undergrad	UK	Consulting	Strategist (Remote)	Competent
P12	35-44 (F)	Undergrad	UK	Energy/Utilities	Head of Operations (Hybrid)	Competent
P13	18-24 (M)	Grad	UK	Education	Research Assistant (Hybrid)	Competent
P14	25-34 (F)	Postgrad	UK	Tech/IT	UX Consultant (Remote)	Expert
P15	25-34 (M)	Undergrad	UK	Tech/IT	Software Engineer (Remote)	Novice
P16	25-34 (M)	Postgrad	UK	Education	Researcher (Hybrid)	Expert
P17	25-34 (F)	Undergrad	UK	Consulting	Lawyer (Remote)	Novice
P18	35-44 (M)	Grad	UK	Tech/IT	Proposition Manager (Hybrid)	Competent
P19	25-34 (M)	Postgrad	UK	Tech/IT	Software Engineer (Hybrid)	Expert
P20	18-24 (M)	Undergrad	UK	Tech/IT	Consultant (Remote)	Novice

Table 1: Participants Demographic Information.

remote work security behaviour. We adopted this approach to help overcome potential reluctance from participants who might be concerned about the consequences of answering such questions honestly or giving answers that are regarded as socially undesirable [7] (i.e. under-reporting undesirable behaviours such as workarounds or non-compliant security behaviour).

The interview was designed to begin by asking general questions about the participant's background, job responsibilities, and remote work experience. Then, questions moved on to security. Participants were asked about whether their remote work has any security or privacy implications, security policies, awareness of security measures, and adherence to security policies, as well as security training for remote work. Lastly, participants were asked questions about their experience with incident reporting and views on security culture at their organisation. Our interview questions can be found in Appendix B.

Prior to each interview, participants were provided with a study information sheet and asked to sign a consent form if they agreed to participate. Subsequently, they completed a

demographic information questionnaire. The interviews were conducted virtually by one of the researchers via Zoom or Microsoft Teams, based on the participant's preference. All interviews were audio-recorded, transcribed, and anonymised. The study exclusively recruited volunteers, who were free to withdraw at any time and for any reason, and no compensation was provided to participants.

3.3 Pilot Study

Prior to conducting the main study, we carried out a pilot study to test our semi-structured interview script with 3 researchers from our institution who have experience with remote work. [64]. The pilot study helped to ensure the clarity of questions and to identify any issues, limitations, or other weaknesses in the interview script beforehand [41].

Based on the pilot study results, we were better informed of the average duration of our interviews at 51 minutes. Moreover, further refinements were made by identifying sensitive questions where participants might be concerned about the consequences of answering honestly or might give answers that are perceived as socially undesirable (i.e. breaching security policy). By rephrasing those sensitive questions as indirect questions [21], participants could then answer from the perspective of another person. This method was found to be effective at minimising social desirability bias [7]. The pilot interviews were not included in the analysis of the research study.

3.4 Data Analysis

Following the Constructivist Grounded Theory procedure of systematically collecting, coding, analysing and theoretically categorising data [13,63], the conducted interviews were audio-recorded, transcribed, and anonymised by the primary researcher. Then, we analysed the interview transcripts using Nvivo, a qualitative data analysis software. The primary researcher and a second researcher iteratively performed open coding by analysing each interview line by line in accordance with the Constructivist Grounded Theory approach [63], and

compared the new codes to the growing collection of codes (i.e., constant comparison). Researchers met with the principal investigator regularly during the analysis to discuss and refine the identified codes, then shifted more toward categorising codes (i.e., focused coding). We established links among different codes, based on an intense analysis focused on observing the categories and their interconnections. We began theoretical coding by iteratively rearranging our categories until they stabilised and confirming the connections built among them. The researchers generated a codebook of 217 codes.

Data saturation [16, 28, 55] was observed between the 18th and 20th interviews in which no significant new codes emerged from those interviews, and we stopped interviewing. In total, the study material analysed consisted of 16 hours and 58 minutes of recorded interviews (~81,420 words), each on average 52 minutes long (~4,771 words).

To verify the credibility of the codebook, the third researcher cross checked the codes against the interview transcripts. Additionally, we tested for inter-rater reliability and found that the average Cohen's kappa coefficient for all codes was 0.85, which is over 0.80 indicating strong agreement [46]. We also assessed the reliability and credibility of the findings through a complementary triangulation method, specifically member checking [35], in which we randomly selected three participants and asked for their feedback on our findings. All participants confirmed the identified categories and themes, without providing any comments that would introduce new themes. The Codebook is available in Appendix C.

3.5 Research Ethics

Our institution's research ethics committee reviewed and approved the study. A study information sheet, along with a consent form, was presented to participants prior to each interview. This sheet explained the purpose of the study and how the collected data would be handled. Each participant confirmed that they had understood the information provided and agreed to participate by filling out a consent form, retaining the right to withdraw from the study at any time. No participants withdrew from the study. All interview transcripts were completely anonymised and stored securely.

3.6 Limitations

Our study has some limitations common to qualitative research: First, our qualitative study is limited by our sample size and diversity. According to prior work recommendations [13], we interviewed between 12 and 20 remote workers until no significant new codes emerged. Furthermore, we recruited a diverse group of participants from different industries and job roles to increase the likelihood of at least one participant mentioning relevant findings. However, it is important to note that our sample is relatively young. Additionally, our qualitative study seeks to explain and understand a phenomenon rather than surveying or generalising from a sample.

Second, researchers' skills and personal biases can influence qualitative research quality [40]. To overcome this limitation, the primary researcher who conducted all interviews was trained in designing and conducting interviews, since the quality of the questions asked [8] and the skill of the interviewer [36] determine the depth of the data collected.

Third, our study is based on interviews where participants self-reported their own behaviour, and it is common to have social desirability bias in self-reporting studies [7]. To minimise social desirability bias, open-ended and indirect questions were used instead of leading questions and participants were encouraged to provide in-depth answers in their own words.

Fourth, a limitation of our study is the potential discrepancy between participants' beliefs about their organisation's security policies and the actual policies in place. Participants may misunderstand official policies due to factors such as poor wording, incomplete knowledge, or changes in policy over time. While we acknowledge this limitation, the self-reported views of participants remain relevant to understanding the motivations behind their actions. Future research may benefit from strategies aimed at validating participants' perceptions against documented security policies.

4 Results

In this section, we present the findings of our study and discuss our key findings organized according to the main themes of our analysis, noting that no significant differences were observed between fully remote and hybrid employees during the analysis. The main themes are: Personal Security Model (Section 4.1), External Security Influences (Section 4.2), Organisational Security Model (Section 4.3), and Personal-Organisational Security Appraisal in Remote Work (Section 4.4).

4.1 Personal Security Model

A Personal Security Model is one of the dominant emergent themes from our study. It is composed of an individual's attitude, perception, knowledge, concerns, beliefs and practices related to personal security. Our data analysis showed that this model is constantly shaped and influenced by an individual's experiences and interactions with their environment, as illustrated in Section 4.2. Furthermore, it guides their personal behaviour in safeguarding both their home and remote work security.

Based on our findings, participants whose personal security models are focussed on productivity regard security as a lesser priority, while participants whose personal security models are aligned with strong security beliefs will prioritise proactive security practices regardless of what is stated in the policy. P14 who works as UX consultant with a productivity mindset said "*It's in this day and age where all that you are forced to think about is hustling and productivity and kind of producing, producing, producing every day. Security takes*

a back step, you would not mind ignoring security rules if it means that you can get things done faster if it will help you that day, if it will help you for the next 5 minutes.". Our analysis identifies the following sub-themes within the personal security model: proactive personal security practices (Section 4.1.1) and faulty security practices (Section 4.1.2).

4.1.1 Proactive Personal Security Practices

Several participants mentioned proactive security practices for protecting either their work-related or personal online activities at home. These included, but not limited to, rules of thumb for checking email legitimacy (i.e., checking email headers, looking up unknown email addresses on Google, scrutinising email content), website authentication (i.e., accessing websites from bookmarks, checking security certificates, inspecting website URLs), and installing new software for both work and personal use (i.e., installing software from the original or trusted source, testing untrusted software on a dedicated machine). P03 explained testing new software on a dedicated machine "*what I used to do is exactly before I installed it and started using it, I used to test it on a different machine just to understand clearly what it was doing and then see what it was doing in the background as well and then start using it.*" P16, who works as a researcher, mentioned checking email legitimacy: "*I think I'm more cautious than others because I'm usually validating the e-mail headers.*" While P17, a lawyer at a consulting company, stated: "*I probably just copy paste the e-mail into Google and just check if it's legitimate or spam.*"

A timely response to security updates was mentioned as a practice by P18, a proposition manager, who noted: "*I just do it because that's what you're supposed to do. I don't know fundamentally why, but I just know because it's cybersecurity. Whatever security patch exists now, they're going to figure it out. It will be a vulnerability that appears at some point, so they detect it, and they create a patch you have to download.*". Driven by their personal privacy concern, P15, a software engineer, mentioned the practice of using separate browsers for work and personal use. P04, a Product Manager, mentioned using complex passwords and changing them frequently as a personal security practice, even though it is not mandated by their startup company. They said, "*I make sure to have complex passwords and change them every so often even though I am not asked by my company.*" Other participants mentioned using personal MFA (N=3), VPN (N=3), and a password manager (N=2).

4.1.2 Faulty Security Practices

Some participants reported faulty security practices stemming from misconceptions or incorrect beliefs. For instance, P17 perceived public WiFi in reputable places as secure, which led them to connect without a VPN. They said, "*I try to go to places that are reputable like Starbucks or those kinds of coffee shops that are chains, and I know they have probably*

got good, secure WiFi in place for their customers." Additionally, P08 conveyed another false perception about the safety of public WiFi of the hotel or cafeteria, stating: "*...in my opinion, these places just want to cater to people's needs, which is WiFi. I don't think they have the intention to steal people's data or whatever.*" However, it is worth noting that public or open WiFi networks are often unsecured and can be vulnerable to malicious attacks such as 'Evil Twin' attacks [59], making them an easy target for hackers looking to steal data. Both participants mentioned the existence of policies that restrict the use of public WiFi for work. Therefore, good security practices advise using a secure WiFi network or VPN when connecting to public networks.

4.2 External Security Influences

A set of influencing factors on the personal security model was identified during the analysis, as depicted in Figure 5. This model of external security influences plays an important role in shaping aspects of the personal security model, including knowledge, attitudes, concerns, and beliefs. These factors consequently affect the personal security decision-making process for both work and non-work contexts. These influences stem from various sources. The diverse nature of influences on the personal security model, in terms of how and from where individuals are influenced, alters the type of influence. For instance, while knowledge serves as a fundamental influencer, providing individuals with the necessary information to assess risks and adopt protective measures, skills represent a distinct category of influence. Skills encompass the practical abilities individuals possess to implement security practices effectively. This could include proficiency in using security tools or navigating digital environments securely.

The identified sources of influence are: online services and technologies, qualifications, external advice, past incidents, past work experiences, and the current organisational security model. In the remainder of this section, we will describe each source of influence.

External advice is sought by individuals such as P18, a Software Engineer, who seeks guidance from a friend skilled in security, complementing their novice competency in security skills. Additionally, others, like P06, an Associate Software Consultant, noted seeking advice from experienced co-workers or IT staff members. P06 expressed, "*I would just pretty much piggyback on everything that more experienced people have done.*"

Past incidents, such as the breach involving unauthorised access to patient information, as reported by P15, and an incident where P10's colleague's laptop was stolen from their car, underscore the importance of protecting personally identifiable information and work devices for them. Additionally, fraudulent banking transactions experienced by P09 led them to close all tabs when accessing their bank accounts and sometimes only use the bank-associated app, as additional measures they take to enhance security and minimise risks.
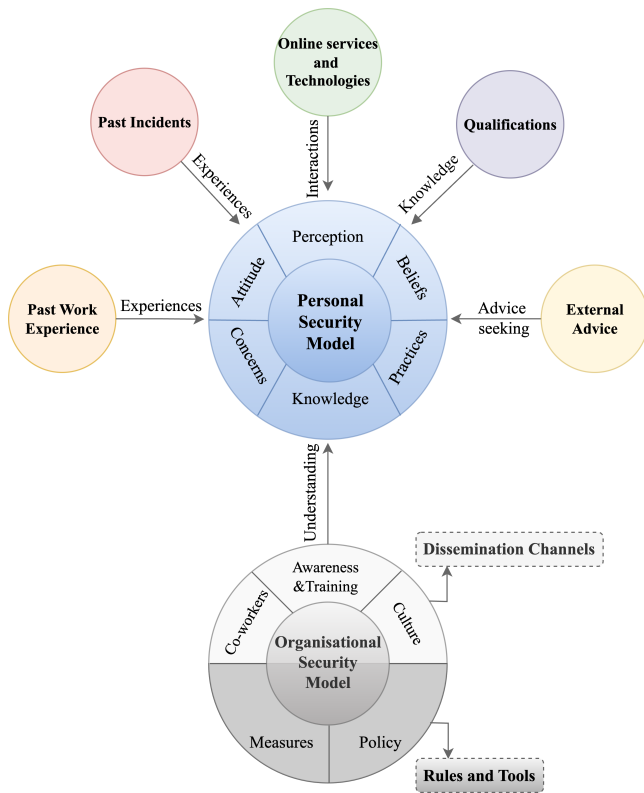
Figure 2: A model of external security influences.

These events serve as valuable lessons that influence the personal security model. In particular, the first two cases emphasise the significance of being vigilant and proactive in safeguarding data and devices to mitigate potential risks and protect oneself from future security breaches.

Online service and technologies, online services encompass various categories, including social media platforms, cloud computing services, financial services, and more. Each category provides a unique user experience and implements distinct security policies and measures. Our analysis revealed evidence suggesting that users can be influenced by their interactions with any type of online service. For example, P20, a Client Solutions Manager, emphasised the impact of encountering policies such as password complexity frequently on their practice in creating personal account passwords. Furthermore, our findings suggest that individuals' perceptions of technology security are often shaped by their interactions with and the popularity of these technologies. P09, for instance, expressed a preference for Apple products, citing their strong reputation for security and consistent security patches, as well as their user-friendly prompts for updates. This consistent approach has significantly influenced P09's attitude toward purchasing their products and installing security updates, despite their limited technical understanding.

Qualifications, our analysis revealed a multifaceted aspect to this source of influence, encompassing differing socioeconomic statuses as a factor alongside educational backgrounds

that range from the quality of education to technical specialty, and their relationship to security. Additionally, other factors include digital access to technology, which shapes individuals' personal security models for action in specific situations. For instance, P12 noted that the great job opportunities they experienced strongly shaped their security-related behaviour. P03, a Security Professional by degree, also commented on their practice of testing unknown tools in different virtual machines and related that to their skills and educational background. Moreover, P01's skill in using video editing tools helped them in cropping identifiable elements of patient video while working remotely with a research partner who has a strict policy regarding data privacy.

Past work experience is a source of influence that shares similar characteristics with organisational security influences. Individuals' personal security models are shaped by the skills and knowledge they acquired through their past work experiences, which can manifest as security practices for personal matters, such as adopting a personal password manager, as noted by P04, as well as for managing their work accounts' passwords with their current employer, who does not provide a password manager.

Organisational security model is the final source of influence on the personal security model, and represents how the current organisation tackles remote work security: both by communicating what employees should be doing, and by providing security rules and controls for them to implement (see Section 4.3). The organisational security model relates to the personal security model in a number of different ways. described in more detail in Section 4.4.

4.3 Organisational Security Model

The organisational security model is another emerging theme from our data analysis. Most organisations define security through a combination of rules and tools (i.e., security policies and security measures) that describe what individuals should and should not do, and provide them with the technical means of doing so (e.g., VPN, endpoint management, MFA). These rules and tools are communicated to remote workers by direct and indirect dissemination channels (i.e., security awareness and training, security culture and co-workers). Security culture is defined as a set of collective norms and values, developed through employee interaction with security elements or experience of the behaviour of their colleagues [17, 62].

As illustrated in Figure 5 , individuals develop a personal understanding of security rules and tools. This understanding is significantly shaped by formal initiatives implemented for disseminating information about these rules and tools, such as security awareness and training programmes. However, personal understanding is also influenced indirectly by co-worker dynamics, organisational security culture, and their personal background. Within this theme, our analysis captured how participants relate to elements of the organisational security model for remote work, as will be illustrated in the

following two sub-themes: tools and rules (Section 4.3.1) and dissemination channels (Section 4.3.2).

4.3.1   Rules and Tools

Security Policies: participants reported different perceptions and attitudes towards security policies, ranging from a lack of clear security policy for remote work. P12, when asked about their familiarity with the policy and guidelines for remote work, said "*there is nothing specifically for remote work.*" P03 confirmed that, "*...most companies do not have a policy. They are just sending emails, giving you guidelines. I don't think they developed policies per se.*" On the other hand, with the existence of policies, P05 mentioned accessibility issues related to policy content, saying "*policies are written in such a way that no one wants to read them because they're written in kind of legal jargon, and no one wants to read through 10 pages of legal jargon just to be told that you shouldn't visit bad websites.*" While P14 remarked, "*..these rules are not for everyday people, it's for computer scientists.*"

Participants (N=6) expressed difficulty in remembering the policies. For example, when asked about their familiarity with the security policy and guidelines provided by their organisation for remote work, P04 responded, "*I don't fully remember what it says.*" In addition, participants (N=5) commented on the lack of policy flexibility. P13 expressed, "*I think it is just done more as a blanket, everyone this is the security; this is the restrictions you will have; you are not allowed to download anything, whereas I think it needs to be done on a more specialised basis.*" Meanwhile, P07 referred to the policy as one-size-fits-all.

Participants were asked about what motivates them to follow policy rules. P05 prioritised job performance in terms of efficiency and effectiveness, they commented, "*I think if the policies match how I need to do my job or make my job easier and protect it.*" While privacy concerns were the driver for policy adherence for P8, a Full-Stack Developer. When asked about what motivates individuals to follow policy rules, they said, "*...as long as they can work productively and not be tracked.*"

Security Measures, along with the provided software and hardware for remote work, are essential for upholding adherence to remote work security policies. Participants have varied understanding, perceptions, and attitudes towards remote work facilities and the security measures in place. Some participants perceived the security measures for remote work as heightened (N=6), where the complexity of security protocols can sometimes clash with practical work demands, prompting the adoption of workarounds. Based on participants' statements when asked about the motivations behind adopting workarounds, P06 mentioned, "*... definitely comfort. Honestly, it's because the procedures are very painful.*" And P7 stated, "*the fact that if something is still too difficult, people will find another way that's probably outside policy to make things happen.*"

Additionally, P13 commented on the contrast between heightened security measures in remote work and office settings, suggesting,"*I just think it's because the hardware they give you to try to be more secure because they know you're not in the office space.*" Furthermore, P15, a Software Engineer at a startup company, highlighted the absence of proactive security measures, pointing out a tendency to neglect certain security aspects under the assumption that negative events will not occur while working remotely.

4.3.2   Dissemination Channels

Security Awareness and Training Programs are considered key components of organisational security initiatives, providing essential knowledge and skills to enhance overall security. Conceptually, this aims to influence the knowledge, practices, and concerns of participants to improve their competence and awareness and to align their concerns with those of the organisation (see Figure 5).

Participants have varied attitudes and perceptions toward the security training provided by their organisation. Participants have reported a lack of quality content (N=3), fatigue from training duration (N=3), repetitive training material (N=4), and questioning the necessity to repeat the same training again and again, resulting in a lack of training efficacy. Using aeroplane safety announcements as an analogy, P15 explained that repetition of basic training content decreases attention and engagement. Other participants reported the lack of comprehensive formal security training (N=5). P08 said, "*We haven't really received any sort of security training.*"

Furthermore, a number of participants proposed ideas to improve the efficacy of the training (N=7). P03 suggested that the training should be chunked, focused, and theme-based training sessions. Moreover, their expertise as a security professional enabled them to recognize specific instances that could impact the security practices of others. P03 proposed utilising hypothetical security scenarios as a means to educate employees.While P18 suggested signposting the new training content so workers are aware of what is new and different from the previous training, which would increase their attention and enhance their learning experience. Also, P18 suggested that security training should be customised based on the worker's background and experience, taking into account their familiarity with previous training and relevant knowledge.

The frequency of the training was discussed by several participants (N=5), with participants proposing monthly, bimonthly, or every four months as a suitable frequency. P07 mentioned that, "*basically anything you do less than quarterly in terms of training will be forgotten.*" P06 suggested, "*Something like an hour every 3 or 4 times a year that would be helpful.*" While P03 suggested 15 minutes training that is very well focused and to be done monthly or bimonthly.

Security Culture and Co-workers act as indirect channels through which employees perceive the security rules and

tools, consequently impacting the overall security posture of the organisation. The absence of immediate in-person support while working remotely can significantly impact how employees approach security. P06 highlighted this by saying, "*I think you are a bit more self-reliant when you are on your own. In theory it's the same as in the office you can always reach someone on the company's chat and then you would get help. That's the theory, right? And in practice, you're more on your own when you're working alone, and you try to do workarounds that you wouldn't necessarily try on your own if you were in the office.*" This sentiment underscores the importance of fostering a supportive security culture, especially in remote work settings, where employees may feel isolated and more inclined to find insecure shortcuts to complete their tasks.

Moreover, interaction with co-workers has multiple influences, which could have a positive or negative outcome. One example noted by P18 is the use of WhatsApp by their co-workers to share work documents as an informal communication channel, ignoring the policy rule prohibiting it. As stated by P18, their behaviour was influenced by interactions with other co-workers, leading them to use unauthorised communication channels for work.

4.4 Personal-Organisational Security Appraisal in Remote Work

Our analysis has shown that user security-related behaviour in remote work is influenced by an appraisal process, as depicted in Figure 3. This process occurs between the users' personal security model and their understanding of the organisational security model rules and tools. The understandings of the rules and tools are gained through dissemination channels, collectively forming the organisational security model, as explained in Section 4.3.

We captured various types of alignments between the personal security model and the organisational security model, characterised by the size and extent of their overlap. These alignments have been summarised into three representative models of alignments, as illustrated in Figure 4. These three models reflect how remote workers subjectively understand and interact with the rules and tools of the organisational security model for remote work. Therefore, our assessment of reported security-related behaviour is not grounded in objective truth, but rather in participants' justification and interpretation of these elements.

4.4.1 Personal and Organisational Security Models are Well Aligned (Figure 4A)

This case represents an well integrated situation where users perceive no limitations in the provision of remote work facilities (i.e., software, hardware, security policies, and measures). Participants reported compliant behaviours with security policies (e.g., performing work tasks on organisation-provided

devices, refraining from USB usage, using VPN, using recommended tools only, using multi-factor authentication, and using complex passwords). P07 pointed out that the satisfaction of all their needs motivated them to follow company security policies to perform work tasks on organisation-provided devices, stating "*...the hardware I have been given is very powerful and easily does all of the things I need to do. So from that perspective I do not need to look for other devices...*"

4.4.2 Personal and Organisational Security Models are Partially Aligned (Figure 4B)

In this case, the two models are partially aligned, where users are mindful of security to varying extents based on both their personal security understanding and on their perception of organisational security. This led to the emergence of three distinct behavioural patterns: poorly compliant security behaviours, proactive security behaviours in the absence of policy, and non-compliant security behaviours driven by security. Notably, the latter two behaviours are instances of shadow security, where users may resort to their own methods of ensuring security, either because they perceive gaps in organisational security or because they feel the need to take additional precautions beyond what is officially mandated.

Poorly compliant security behaviours: In this case, the participants do not behave according to the desired security behaviour. Instead, they comply with the policy but disagree with it, leading to less secure behaviour driven by compliance. This included sporadic VPN usage for work and password reuse. Participants discussed the influences behind such poor behaviour. P08, a Full-Stack Developer, stated that personal privacy concerns and VPN drawbacks are the main reasons behind occasional VPN usage for work, saying, "*They provide the VPN from Cisco and it's kind of slow and laggy and I kind of don't like it... Well, they give us the VPN for security, but you know they're in fact monitoring me. So no, I don't really use it on a daily basis. I just use it occasionally.*"

A stated need for convenience and memorability led P19, a Software Engineer, to reuse one password for their work device and the password manager, in addition to laptop login constraints that prevent the use of PIN – a set of numbers – over passwords. P19 said "*...I am reusing one password for logging on to Windows as well as the password manager, they're the same password. Usually I don't do that, but for work I needed a password to remember, and I wasn't going to make more than one...*" Furthermore, P19 added, "*...it's because they're forcing me to use a password, not a PIN on my laptop. I can't log in with the pin. I need the full password. So, I just use the same one I used for my password manager as well. I think it's a strong password.*"

Proactive security behaviours in the absence of policy: In this case, several behaviours aimed at improving participants' remote work security were reported when formal security policies are not in place, including: enhancing home WiFi security (e.g., changing WiFi password regularly, monitoring
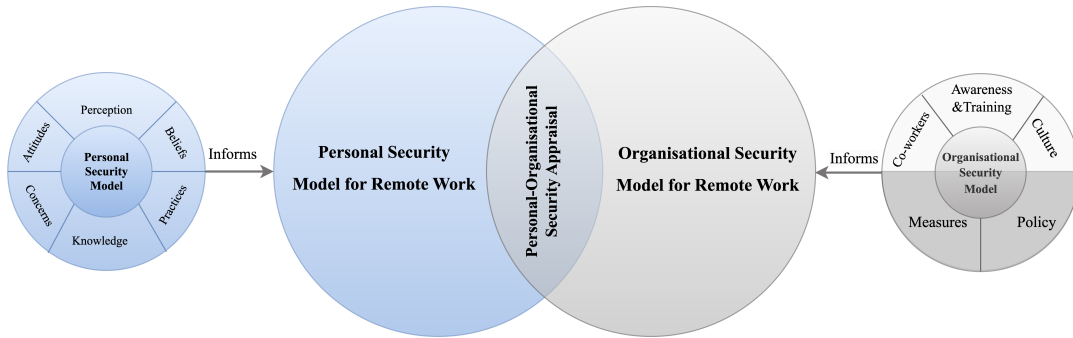
Figure 3: A model of alignment between personal and organisational security models for remote work

Personal Security Model

Organisational Security Model
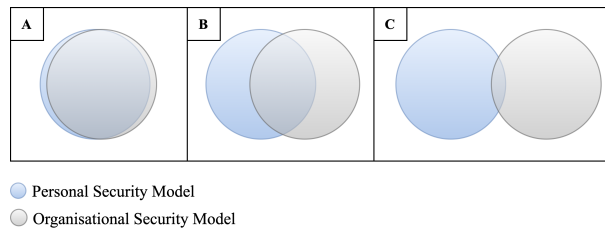
Figure 4: Different modalities of alignments between personal and organisational security models for remote work

and controlling connected devices), daily laptop shutdown, installing software only from trusted sources, segregating work and personal devices, using complex passwords, using a shredder at home, using secure file sharing, and avoiding suspicious websites on work laptops.

Non-compliant security behaviours driven by security, involve users' behaviours that deviate from established security policies but still consider security with alternative means. For instance, P02 mentioned using a secure file sharing platform like Secure Dropbox as an alternative due to limitations with the cloud service provided by their company, despite the policy prohibiting such action. P02 prioritised security and sought out a solution that better met their needs, stating, "*I primarily use Dropbox just because you need to log in and there are some security measures there.*"

4.4.3 Personal and Organisational Security Models are Poorly Aligned (Figure 4C)

In this case, participants reported instances of non-compliant behaviours that could undermine their remote work security. These behaviours are driven by various factors other than their interpretation of security policies. These include connecting to public WiFi without VPN, substituting the recommended software or tools without permission, transferring data between personal and work device, sharing work documents via WhatsApp, using insecure file sharing service (WeTransfer), sharing account passwords with co-workers, and bypassing print restriction by sending work documents to personal email.

All reported behaviours here were perceived by participants to be in breach of an underlying policy rule and mainly

driven by convenience. P05, commented on sending work documents to personal email due to restrictive printing policy that does not align with their work. Admitting the behaviour to be risky, they said "*...there's a lot of restrictions over what can be printed or sent and at the end of the day if someone needs to print something. It means they have to share it to their personal e-mail and then print it. So, that's where the policies don't match the work and the workaround is where the risk is.*" Also, they commented on using insecure file sharing service (WeTransfer) over the company recommended solution (SharePoint),"*...our company is taking a policy that we can't do or download from WeTransfer so that makes it just an extra hassle for people's work... I think consumer solutions like WeTransfer solves that as the easiest case whereas SharePoint there's just so many more extra steps to get what you need done and it's so easy to forget.*"

Other participants reported workarounds driven by productivity such as creating backdoors to access internal resources remotely, replacing the hard drive on the work device, or performing work on personal devices in order to eliminate restrictions. We did not expect our participants to discuss their own personal and deliberate breaches of policy. However, when asked why someone would make use of workarounds in remote work settings, they provided several justifications. These included beliefs about limited organisational monitoring in remote work, human nature preferring ease of use, privacy invasion concerns, slow or relaxed IT response, productivity reasons, and underestimation of the security threat posed by the workaround.

5 Discussion

As traditional organisational boundaries become less tangible, more flexible, and more porous, our results show that shadow security practices continue to evolved to match.

Remote Work Security Policies: Our study highlights that shadow security in remote work encompasses behaviour that aims to improve, extend, or remediate the perceived limitations of existing security policies. A number of these limitations were directly tied to the security policies themselves. The first policy limitation was that some participants could not remember the details of security policies or felt that these policies were not clearly written and communicated. This limitation is relatively straightforward, centred broadly on problems with the timeliness, language and communication of the policies themselves.

A second policy limitation is more subtle and our participants articulated this as policies that were not suited to their needs, leading to frustration, friction or other impediments. These problems arise from an individual's subjective assessment of the security policies, looking at the perceived need, effectiveness, and cost/benefit of following the policy. These findings aligns with previous studies [6, 39] which highlight how shadow security practices can emerge due to perceptions of inappropriate organisational policies and rules. We describe this as the personal-organisation security appraisal, and note that there are commonalities between the personal security model and the Theory of Planned Behaviour (TPB) [4], one of the most widely used theories for studying user attitudes as an influence on human behaviour. TPB defines four factors that underlie the decision toward certain behaviours: attitudes, subjective norms, perceived behavioural control, and intentions. Since shadow security is highly tied to the user's personal security model, which comprises their attitudes, perceptions, knowledge, concerns, beliefs, and practices related to personal security, it encompasses all the elements that can influence a person's decision to behave in a certain manner.

We believe that policy authors, such as CISOs, need to be particularly aware of the content, delivery, and uptake of remote work security policies, as compliance, non-compliance and shadow practices may be harder to determine.

Organisational Security Awareness, Training, and Education (SATE): Our results also suggest that while the personal security model is strongly tied to individual attitudes, perceptions and beliefs, it is also shaped by previous and ongoing SATE efforts. As mentioned in Section 4.3.2, SATE targets individuals to improve their knowledge, upskill their practices, and influence their concerns to be better aligned with the needs of the organisation that employs them. We believe that there are interesting implications arising from the fact that high quality SATE can benefit future employers of existing employees. Put another way: current employers benefit (or suffer) from the SATE efforts of previous employers. Organisations directly benefit from improving the security knowledge and skill of their employees, however there are also positive externalities for other employers who benefit when those trained employees are then recruited. With the rise of the gig economy [37], this has particular implications on the economics, delivery, and alignment of SATE in the context of employees that have multiple employers.

Informal Communications: In tandem with SATE, we also found that remote employees rely on indirect channels to learn and share security know-how with other employees. Our findings suggest that remote workers are more isolated from their peers and the security culture of their employing organisation. This may undermine information sharing between colleagues about security practices and rules, leading to poor understanding of rules and fewer opportunities to learn how to use tools correctly. These informal dissemination channels are much less developed in remote work settings, and our findings indicate this is likely to contribute to poor or non-compliant security behaviour.

Usability of Remote Work Security: Finally, we note that shadow security practices can arise from technical limitations in the provision of remote work facilities. Our participants mentioned that some of the controls they had to use (e.g. access control) were complex and constraining, leading to difficulties in achieving their work objectives. In addition, participants also noted that there was a lack of available support options, meaning they felt more isolated and had to solve problems themselves. Both of these issues are indicative of the need for greater consideration of usability and the wider security user experience for remote workers.

Further research into shadow security practices for remote work can provide a fruitful source of inspiration and innovation, helping to shape new ways of working remotely and securely. Our recommendation are consistent with the approach taken by Kirlappos et al. [39], which aims to learn from shadow practices to improve overall organisational security. As Kirlappos et al. [39] aptly state, "*shadow security existence should not be treated as a problem, but as an opportunity to identify shortfalls in current security implementations that can be leveraged in providing more effective security solutions for organisations.*" By embracing this perspective, organisations can address the gaps in their current security measures and develop more effective and user-friendly security solutions for remote work environments.

6 Recommendations

Based on our findings, we discuss the following recommendations:

6.1 Developing Informal Security Channels

A key finding from our study is the important role of colleagues as a source of security information. An organisation whose employees access shared spaces and communicate

face-to-face can expect informal and private communications to happen spontaneously. However in a remote work environment, such communications need to be a) mediated technologically, b) initiated deliberately, and c) responded to purposefully. One problem arising from a) is that employees feel that communications are more difficult in remote work settings, and we also noted some concerns about companies monitoring their remote employees, both of these concerns can hinder the open discussions about security rules and tools among colleagues. Furthermore, b) and c) both create barriers to spontaneous or opportune discussions that can occur outside of a deliberately initiated interaction. As a result, we argue that remote workers need better technology to help them connect with co-workers about security issues and to share their concerns and solutions, and that more research is needed to determine how and when informal security discussions can be supported to improve security culture among remote workers.

6.2 Homogenising Security Interactions

Individuals are often influenced by their interactions with various platforms such as devices and services, particularly regarding security protocols and practices, which may vary across platforms. This variability can either foster secure habits over time through consistent exposure to the same protocols or lead to confusion and resistance when changes occur, potentially resulting in actions that could pose security vulnerabilities. A key finding from our study is that habit and convenience were among the factors considered during the personal-organisational security appraisal, leading to poor compliance behaviour, shadow security, and even non-compliance with security policies. It is also worth noting that a corollary to this is that innovation and change are particularly difficult in security, as this aims to break previous modes of interaction and familiarity in favour of new ones. Particular attention should therefore be placed on exploring how and when change is necessary, together with suitable strategies for introducing and managing change.

These insights underscore the necessity for standardising security tools and regulations, especially in remote work, which is increasingly prevalent across diverse industries, each facing unique requirements security challenges. To tackle this complexity, we propose implementing security style guides specifically tailored for remote work environments, aiming to homogenise security interactions across platforms and industries. These guides will serve as comprehensive resources outlining best practices, policies, and procedures for ensuring the security of remote work setups. By integrating insights from various industries, security practitioners can develop comprehensive guidelines addressing a wide range of security concerns, fostering knowledge sharing and collaboration across industries.

6.3 Adopting User Experience Design for Remote Work Solutions

Our study identifies that poorly designed remote work solutions can significantly hinder productivity, increase frustration, and elevate security risks. These frustrations often compel employees to create workarounds and shadow security practices. By prioritising user experience (UX) design [12, 45] in the development of remote work solutions, organisations can create intuitive interfaces and streamline workflows that encourage compliance with security measures. UX not only enhances user satisfaction but plays a critical role in ensuring adherence to security protocols. This involves conducting user research, gathering feedback from remote workers, and iteratively refining the design of remote work tools and platforms to prioritise usability and security simultaneously. By adopting a user-centred design approach and aligning user experience with security objectives, organisations can foster a culture of compliance and reduce the prevalence of workarounds and shadow security practices among remote workers.

7 Conclusion

Our exploratory study of security and shadow security practices in the context of remote work was motivated by the prevalence of remote work in the knowledge economy and the lack of research in this context. Based on our analysis of 20 semi-structured interviews with remote workers, our findings complement and extend prior research, which found that shadow security practices can arise from perceptions of inappropriate organisational policies and rules [6, 39].

Our analysis proposes three models for describing security practices in remote work: the first is a personal security model influenced by external factors (e.g. past experiences, knowledge of technology, or qualifications). The second comprises the current organisational security model for remote work, which includes security rules and tools disseminated through awareness and training, interaction with colleagues, and the overall security culture. The third is an appraisal process individuals use to relate the personal and organisational security models, driving compliance, non-compliance, and shadow security behaviour in remote work settings.

This opens up opportunities for future research in remote work security, for example exploring the delivery and long term effects of security awareness, training, and education for remote work in the gig economy; tackling the challenge of improving and harmonising security user experiences across different device and service providers; or exploring how informal communications can be facilitated in remote work settings. It also allows for the investigation of different interventions, such as persuasive techniques or digital behaviour interventions, as a means to enhance user security behaviour in remote work settings.

Acknowledgments

References

[1] Definition of remote work - gartner information technology glossary. https://www.gartner.com/en/information-technology/glossary/remote-work.

[2] 2022 workplace trends & insights report. https://www.beezy.net/2022-workplace-report, January 2023. Retrieved on 2023-01-24.

[3] Anne Adams and Martina Angela Sasse. Users are not the enemy. *Communications of the ACM*, 42(12):40–46, 1999.

[4] Icek Ajzen. The theory of planned behavior. *Organizational behavior and human decision processes*, 50(2):179–211, 1991.

[5] Tammy D Allen, Timothy D Golden, and Kristen M Shockley. How effective is telecommuting? assessing the status of our scientific findings. *Psychological science in the public interest*, 16(2):40–68, 2015.

[6] Adam Beautement, M Angela Sasse, and Mike Wonham. The compliance budget: managing security behaviour in organisations. In *Proceedings of the 2008 new security paradigms workshop*, pages 47–58, 2008.

[7] Nicole Bergen and Ronald Labonté. "everything is perfect, and we have no problems": detecting and limiting social desirability bias in qualitative research. *Qualitative health research*, 30(5):783–792, 2020.

[8] Peter Birmingham and David Wilkinson. *Using research instruments: A guide for researchers*. Routledge, 2003.

[9] Mary Bispham, Sadie Creese, William H Dutton, Patricia Esteve-Gonzalez, and Michael Goldsmith. Cybersecurity in working from home: An exploratory study. In *TPRC49: The 49th Research Conference on Communication, Information and Internet Policy*, 2021.

[10] John M Blythe, Lynne Coventry, and Linda Little. Unpacking security policy compliance: The motivators and barriers of employees' security behaviors. In *Eleventh Symposium On Usable Privacy and Security ({SOUPS} 2015)*, pages 103–122, 2015.

[11] Charles F Cannell, Peter V Miller, and Lois Oksenberg. Research on interviewing techniques. *Sociological methodology*, 12:389–437, 1981.

[12] George Chalhoub, Ivan Flechais, Norbert Nthala, Ruba Abu-Salma, and Elie Tom. Factoring user experience into the security and privacy design of smart home devices: A case study. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–9, 2020.

[13] Kathy Charmaz. *Constructing grounded theory: a practical guide through qualitative analysis*. Introducing qualitative methods. Sage, London, 2006.

[14] Kathy Charmaz. *Constructing grounded theory*. Introducing qualitative methods. SAGE Publications Ltd, London, 2nd edition edition, 2014.

[15] Kathy Charmaz. Constructivist grounded theory. *The Journal of Positive Psychology*, 12(3):299–300, May 2017.

[16] Juliet Corbin and Anselm Strauss. *Basics of Qualitative Research (3rd ed.): Techniques and Procedures for Developing Grounded Theory*. Thousand Oaks, California, 2008.

[17] Adéle Da Veiga and Jan HP Eloff. A framework and assessment instrument for information security culture. *Computers & security*, 29(2):196–207, 2010.

[18] Stuart E Dreyfus and Hubert L Dreyfus. A five-stage model of the mental activities involved in directed skill acquisition, 1980.

[19] Durga Prasad Dube and Rajendra Prasad Mohanty. Application of grounded theory in construction of factors of internal efficiency and external effectiveness of cyber security and developing impact models. *Organizational Cybersecurity Journal: Practice, Process and People*, 3(1):41–70, 2023.

[20] Yogesh K Dwivedi, D Laurie Hughes, Crispin Coombs, Ioanna Constantiou, Yanqing Duan, John S Edwards, Babita Gupta, Banita Lal, Santosh Misra, Prakhar Prashant, et al. Impact of covid-19 pandemic on information management research and practice: Transforming education, work and life. *International journal of information management*, 55:102211, 2020.

[21] Robert J Fisher. Social desirability bias and the validity of indirect questioning. *Journal of consumer research*, 20(2):303–315, 1993.

[22] Steven Furnell and Jayesh Navin Shah. Home working and cyber security–an outbreak of unpreparedness? *Computer fraud & security*, 2020(8):6–12, 2020.

[23] Barney G. Glaser. *Basics of Grounded Theory Analysis*. Sociology Press, Mill Valley, CA, 1992.

[24] Barney G Glaser, Anselem L Strauss, and E Strutzel. The discovery of grounded theory: Strategies for qualitative research new york aldine de gruyter. *GlaserThe Discovery of Grounded Theory: strategies for qualitative research1967*, 1967.

[25] Timothy Godlove. Examination of the factors that influence teleworkers' willingness to comply with information security guidelines. *Information Security Journal: A Global Perspective*, 21(4):216–229, 2012.

[26] Leo A Goodman. Snowball sampling. *The annals of mathematical statistics*, pages 148–170, 1961.

[27] Lynda Gratton. How to do hybrid right. *Harvard Business Review*, 99(3):66–74, 2021.

[28] Greg Guest, Arwen Bunce, and Laura Johnson. How many interviews are enough? an experiment with data saturation and variability. *Field methods*, 18(1):59–82, 2006.

[29] Ken H Guo. Security-related behavior in using information systems in the workplace: A review and synthesis. *Computers & Security*, 32:242–251, 2013.

[30] Andreas Györy, Anne Cleven, Falk Uebernickel, and Walter Brenner. Exploring the shadows: It governance approaches to user-driven innovation. In *ECIS 2012 - Proceedings of the 20th European Conference on Information Systems*, 2012.

[31] Steffi Haag and Andreas Eckhardt. Shadow it. *Business & Information Systems Engineering*, 59:469–473, 2017.

[32] Stephen Hansen, Peter John Lambert, Nicholas Bloom, Steven J Davis, Raffaella Sadun, and Bledi Taska. Remote work across jobs, companies, and space. Technical report, National Bureau of Economic Research, 2023.

[33] Karin Höne and Jan H. P. Eloff. Information security policy—what do international information security standards say? *Computers & security*, 21(5):402–409, 2002.

[34] Allen C. Johnston, Barbara Wech, Eric Jack, and Micah Beavers. Reigning in the remote employee: Applying social learning theory to explain information security policy compliance attitudes. In *16th Americas Conference on Information Systems 2010, AMCIS 2010*, volume 3, pages 2217–2230, 2010.

[35] Karsten Jonsen and Karen A Jehn. Using triangulation to validate themes in qualitative studies. *Qualitative research in organizations and management: an international journal*, 4(2):123–150, 2009.

[36] Annabel Kajornboon. Using interviews as research instruments. *E-journal for Research Teachers 2.1*, page 1–9, 2005.

[37] Otto Kässi and Vili Lehdonvirta. Online labour index: Measuring the online gig economy for policy and research. *Technological forecasting and social change*, 137:241–248, 2018.

[38] Iacovos Kirlappos, Simon Parkin, and M Angela Sasse. Learning from "shadow security". In *NDSS Workshop on Usable Security*, 2014.

[39] Iacovos Kirlappos, Simon Parkin, and M Angela Sasse. " shadow security" as a tool for the learning organization. *Acm Sigcas Computers and Society*, 45(1):29–37, 2015.

[40] Benjamin Koskei and Catherine Simiyu. Role of interviews, observation, pitfalls and ethical issues in qualitative research methods. *Journal of Educational Policy and Entrepreneurial Research*, 2(3):108–117, 2015.

[41] Steinar Kvale and S Brinkmann. Introduction to interview research. *Doing interviews*, pages 2–11, 2007.

[42] Ying Li and Mikko Siponen. A call for research on home users' information security behaviour. In *PACIS 2011 Proceedings*, page 112, 2011.

[43] Florian Malecki. Overcoming the security risks of remote working. *Computer fraud & security*, 2020(7):10–12, 2020.

[44] Gabriela Labres Mallmann and Antonio Carlos Gastaud Maçada. Behavioral drivers behind shadow it and its outcomes in terms of individual performance. In *AMCIS 2016: Surfing the IT Innovation Wave - 22nd Americas Conference on Information Systems*, 2016.

[45] Aaron Marcus. *HCI and user-experience design*. Springer, 2015.

[46] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012.

[47] Ferdinando Monte, Charly Porcher, and Esteban Rossi-Hansberg. Remote work and city structure. *American Economic Review*, 113(4):939–981, 2023.

[48] Albine Moser and Irene Korstjens. Series: Practical guidance to qualitative research. part 3: Sampling, data collection and analysis. *European journal of general practice*, 24(1):9–18, 2018.

[49] JM Nilles, FR Carlson, Paul Gray, and Gerhard Hanneman. Telecommunications-transportation tradeoffs. *Final report*, 1974.

[50] Norbert Nthala and Ivan Flechais. Informal Support Networks: an investigation into Home Data Security Practices. pages 63–82, 2018.

[51] Savvas Papagiannidis, Jonathan Harris, and David Morton. Who led the digital transformation of your company? a reflection of it related challenges during the pandemic. *International journal of information management*, 55:102166, 2020.

[52] Lazar Raković, Marton Sakal, Predrag Matković, and Mirjana Marić. Shadow it–systematic literature review. *Information Technology and Control*, 49(1):144–160, 2020.

[53] Jason Sabin. The future of security in a remote-work environment. *Network Security*, 2021(10):15–17, 2021.

[54] Martina Angela Sasse, Debi Ashenden, D. Lawrence, L. Coles-Kemp, I. Fléchais, and P. Kearney. Human vulnerabilities in security systems. *Human Factors Working Group, Cyber Security KTN Human Factors White Paper*, 2007.

[55] Clive Seale. The quality of qualitative research. *The Quality of Qualitative Research*, pages 1–224, 1999.

[56] Mario Silic and Andrea Back. Shadow it–a view from behind the curtain. *Computers & Security*, 45:274–283, 2014.

[57] Mario Silic, Jordan B Barlow, and Andrea Back. A new perspective on neutralization and deterrence: Predicting shadow it usage. *Information & management*, 54(8):1023–1037, 2017.

[58] Darja Smite, Nils Brede Moe, Jarle Hildrum, Javier Gonzalez-Huerta, and Daniel Mendez. Work-from-home is here to stay: Call for flexibility in post-pandemic work policies. *Journal of Systems and Software*, 195:111552, 2023.

[59] Yimin Song, Chao Yang, and Guofei Gu. Who is peeping at your passwords at starbucks?—to catch an evil twin access point. In *2010 IEEE/IFIP International Conference on Dependable Systems & Networks (DSN)*, pages 323–332. IEEE, 2010.

[60] VA Stafford. Zero trust architecture. *NIST special publication*, 800:207, 2020.

[61] Anselm Strauss and Juliet Corbin. *Basics of Qualitative Research Techniques*. Sage Publications, Thousand Oaks, CA, 1998.

[62] Kerry-Lynn Thomson, Rossouw Von Solms, and Lynette Louw. Cultivating an organizational information security culture. *Computer fraud & security*, 2006(10):7–11, 2006.

[63] Robert Thornberg, Kathy Charmaz, et al. Grounded theory and theoretical coding. *The SAGE handbook of qualitative data analysis*, 5(2014):153–69, 2014.

[64] III Turner, Daniel W. Qualitative interview design: a practical guide for novice investigators. *Qualitative report*, 15(3):754–, 2010.

A   Demographics Questionnaire

1. Select your age group:

   ○ 18-24
   ○ 25-34
   ○ 35-44
   ○ 45-54
   ○ 55-64
   ○ 65-74
   ○ 75 or older
   ○ Prefer not to answer

2. Select your gender:

   ○ Male
   ○ Female
   ○ Other
   ○ Prefer not to answer

3. Where do you live?

   ○ —
   ○ Prefer not to answer

4. What is your work setting?

   ○ Remote: Fully remote work.
   ○ Hybrid: A combination of remote work and working from a designated office space.

5. Which of the following best describes your organisation's business domain?

   ○ Manufacturing
   ○ Retail
   ○ Technology/IT
   ○ Healthcare
   ○ Finance
   ○ Hospitality
   ○ Education
   ○ Consulting
   ○ Real Estate
   ○ Transportation and Logistics
   ○ Entertainment and Media
   ○ Non-profit/NGO
   ○ Government/Public Sector
   ○ Energy and Utilities
   ○ Other —

6. What best describes your role within your organisation?
7. What is the highest level of school you have completed?

   ○ No schooling completed
   ○ Nursery
   ○ High School
   ○ Trade/technical/vocational training
   ○ Undergraduate studies
   ○ Graduate studies
   ○ Postgraduate studies

8. How would you rate your technical skills in computer security and privacy (e.g. understanding threats, vulnerabilities, and countermeasures)?

   ○ Novice
   ○ Competent
   ○ Expert

B Interview Questions

B.1 Remote Work Experience

1. Can you tell us a bit about yourself and your background?
2. Can you tell me about your experience working remotely?
3. How long have you been working remotely?
4. Can you share any specific examples of remote work tasks you have successfully completed in the past?
5. Did you work remotely from home before the pandemic?

    (a) If yes, how frequent?

6. How does your current remote work differ from your previous experiences before the pandemic?
7. Have you faced any challenges while working remotely?

    (a) If so, how did you overcome them?

B.2 Introductory to Security in Remote Work

1. How does cybersecurity fit into your day?
2. Do you think your remote work has any security or privacy implications?

    (a) If yes, what would be your concerns? (Prompt: dealing with confidential information)

B.3 Security, Awareness and Training

1. How familiar are you with the security policies and guidelines provided by your organisation for remote work?
2. Have you received any security training recently?

    (a) If yes, how long ago?
    (b) What was it like? (Prompt: Training format, sessions length)
    (c) Do you think it is helpful?

3. Do you receive reminders about security? (Prompt: Emails, nudges)

    (a) If yes, what do they ask/prompt you to do?

B.4 Personal vs Work Protection

1. Is there anything you do at home to protect remote work over and above what you would normally do for other online activities at home? (Prompt: securing your home WiFi network or using a VPN to access remote resources)

B.5 Remote Work Setup (Equipments and Tools)

1. What devices do you use for remote work?
2. Are they your personal devices or provided by your organisation?
3. Are you ever worried about the possibility of them being lost or stolen?
4. Do you have a routine for regular backups?
5. Are there any specific communication or productivity tools recommended by your organisation?

    (a) If yes, what are they?
    (b) Are they good enough?

6. Do you use other tools?

B.6 Security Policy and Measures

1. Do you think it is important to keep your device and software up to date with the latest security patches and updates?
2. Have you ever installed any software other than that provided by the organisation?

    (a) If yes, why do you do that?
    (b) Did you take any precautions when doing so? (Prompt: verify the source)

3. How do you verify the authenticity of websites or online resources before providing sensitive information, such as login credentials or personal data, while working remotely?
4. What measures do you take to prevent unauthorized access to your remote work device? (Prompt: strong passwords, two-factor authentication, or biometric authentication)
5. Where do you usually perform your job when working remotely? (Prompt: public areas like cafes, at home office)

    (a) If public areas, do you think using public WiFi might pose a threat to the organisation? and how?

6. Do you handle any physical paperwork or print out information related to your work?

    (a) If yes, does any of it include potentially confidential information?
    (b) If it does, how do you dispose of such documents once you're finished with them?

7. Do you share the devices you work on with anyone else in your household?

    (a) If yes, do you believe that this could pose a security threat?
    (b) How do you ensure the protection of your work-related materials?

8. Do you use removable storage devices, such as USB sticks, to store or transfer work-related data?

    (a) If yes, how important is it? Why?
    (b) Is that your own one or was it given to you by the organisation?
    (c) Is any of the stored data in any sense confidential?
    (d) What precautions do you take to protect that data?

9. Is there any situation where you encounter difficulties accessing legitimate resources or platforms?

    (a) If yes, have you ever used a workaround to bypass the restrictions?
    (b) Are you aware if others do the same?
    (c) How frequently does this happen?

B.7 Security Incidents

1. How do you handle unexpected security incidents or potential security threats, such as suspicious emails or notifications, while working remotely?
2. Have you ever come across something that you consider to be a vulnerability that the organisation has not thought of?

B.8 Security Culture

1. To what extent do you believe individuals generally adhere to the policy rules?
2. Can you think of a reason why somebody might not follow one of them?
3. Are there any policies or procedures that you routinely do not comply with? Why do you do this?
4. Does the organisation check whether employees comply with security policies?

    (a) What sanctions or punishments are used against people that get caught?
    (b) Do you think these are appropriate?

5. In general, what do you think of the policies? Do you think they are too strict, too soft, or about right?
6. What is your perception of the overall security culture within the organisation? Would you consider it to be highly security-conscious or not particularly focused on security?

C   CodeBook

Theme	Sub-Theme	Category	List of Codes
Personal Security Model	Proactive Personal Security Practices	Personal Security Practices (general)	Using a separate browser for work and personal use; Using separate work and personal password managers; Timely response to security updates; Using complex passwords; Using DNS over HTTPS; Using DNSSEC; Using a password manager; Using personal MFA; Using a personal VPN; Full disc encryption; Keep fully encrypted backup; Locking device screen when away; Never use public WiFi; No sharing devices in the household; Proactive email verification; Safe browsing practices on work device.
		Rules of thumb website authenticity	Access websites from bookmarks list; Avoid sponsored links in search results; Check the security certificate; HTTPS verification; Inspect the website URL; Look for copyright and trust badges; Look for website reviews; Look up IP address using Whois; Selective trust-based website reputation; Trust Google's first link of the website; Verify using Google Search
		Rules of thumb Email legitimacy	Check email header; Look up email address on Google; Scrutinise email content
		Rules of thumb installing new software (work and personal use)	Avoiding untrusted software installation; Installing software from the original source; Installing software from a trusted source; Installing well-known software only; Testing untrusted software on a dedicated machine
	Faulty Security Practices	Faulty Security Practices	Connecting to reputable (public or any) WiFi; Performing personal activities on work devices.
External Security Influences	Online services and technologies	Online services and technologies	interactions with Online services and technologies; technology access
	Past incidents	Past incidents	past incidents
	Past work experience	Past work experience	security measures; security policies; security training
	External advice	External advice	experienced colleague or friend; online forum; Online search for security insights
	Qualifications	Qualifications	Educational background; Access to job opportunities
	Organisational security model	Organisational security model	Co-workers' practices; Current security reminders; Current security training
Organisational Security Model	Remote work challenges	Remote work challenges	Communication issues with other team members; Connectivity and accessibility issues; Creating structured communication; Device restrictions in remote vs. office work; Fear of losing or damaging work devices; Intangible aspects lost in remote work; Lab equipment accessibility; Lack of clear hybrid-remote work policies; Lack of immediate colleague assistance; Lack of security education; Managing prolonged online discussions or arguments; Mixing personal and work activities on work devices; Resource accessibility issues; Risk from international travel; Theory vs. Practice in remote work assistance; Unofficial vs. official communication; Unstable virtual working environment; Using a public network; Varied remote work locations; Visual exposure of work in public areas
	Rules and Tools	Security measures	perception of enforced software update; perception of immunity to risks; desire for effortless security; equating working from home with a physical office; Balancing security measures and usability; Limited proactive security measures; Endpoint management; MFA; VPN; Modem verification; Zero Trust model; VPN issues
		Security policy (attitudes/perceptions)	Adherence to company policy; Absence of remote work security policy; Adherence to technically enforced policy; Challenges in policy compliance; Challenges in policy content accessibility (language or format); Difficulty in remembering policies; Proactive policy familiarization; Productivity-driven mindset; Risk-driven adherence; Shadow IT and policy adaptation; Demand for tailored security policies; No actions above policy when working at home. Make the work environment unusable; Mixed perceptions about the security policies; No one read the policy; One-size-fits-all policy or Lack of flexibility; Perception of policy disregard; Policies as bureaucratic; Policies as facilitators vs. barriers; Policy and guideline flexibility; Policy Complexity based on business size; Policy is for 'Stupid User'; Positive effect of user-centred security design; Positive view of stricter policies; Privacy-driven adherence to policy; Productivity-driven adherence to policy; Lack of policy on installing new software; Lack of informed incident reporting process; Varied policy implementation by industry
	Dissemination Channels	Security Awareness and Training Programs (attitudes & perceptions)	Importance of ongoing training; Knowledge empowerment; Lack of enthusiasm toward training; Resistance to training; Utility of training (trained vs useful); Effectiveness of incentive-based reminders; Shaping user behaviour; Basic training content; Boredom due to training duration; Distraction from meaningful tasks; Emphasis on policy reinforcement; Enforced security training; Focus on GDPR compliance; Inform the incident management protocol; Lack of belief in training efficacy; Lack of company-specific training; Lack of formal security training; Perceived effectiveness of enforced training; Repetitive security training; Sustaining knowledge through training; Training redundancy over time; Variability in security training based on job role; corrective training
		Security Awareness and Training Programs (suggested enhancements)	Hypothetical security scenarios; Informal learning through discussion forums; Interactive training sessions; Leveraging marketing and PR strategies; Scenarios based on previous incidents; Signposting new content; Tailoring security campaigns to different user types; Tailoring training to worker background and experience; Training frequency; Chunked and focused training session (Theme-based); Utilising communication platforms for security; Cyber score metrics
		Security Culture and Co-workers	Absence of immediate in-person support; Implicit trust in the used software (Start-up case); Lack of support in academic institutions; Prioritising efficiency over security; Reduced contextual awareness when working remotely; Lack of guidance and oversight
Personal-Organisational Security Appraisal in Remote Work	Well alignment	Complaint behaviours	Document labelling or classification; Following policies for installing new software; Maintaining secured home workspace; No sharing of work devices; Performing work tasks on company provided devices; Refraining from USB usage; Timely response to software updates; Updating passwords frequently; Using complex passwords; Using MFA; Using a password manager; Using recommended tools; Using thumbprint USB; Using VPN
		Justifications	Seamless remote infrastructure; Lack of necessities
	Partial alignment	Poor complaint behaviours	Occasional VPN usage for work; Password reuse
		Justifications	Faulty beliefs; Ease of use; Lack of policy; Lack of measures
		Proactive security behaviours no policy	Avoiding suspicious websites on work laptops; Daily laptop shutdown; Enhanced home WiFi security; Installing software from a trusted source; Minimizing visibility to others in public areas; Monitoring home network-connected devices; Never working or connecting to public WiFi; Recycling passwords; Segregating internal and external file sharing; Segregating work and personal devices; Strict work laptop usage; Using complex passwords; Using a hotspot through a mobile data plan; Using a password manager for work accounts; Using personal MFA; Using secure file sharing; Using a shredder at home; Using VPN
		Non-complaint behaviours driven by security	Replace official file sharing with secure alternative
	Poor alignment	Non-complaint behaviours (workaround)	Using insecure file sharing service; Disconnecting from the VPN; Installing new tools or software; Sharing account password with co-workers; Substituting the recommended software; Transferring data between personal and work devices; Creating a backdoor; Replacing work laptop hard drive; Using Google Sheets instead of Excel; Using personal email for work; Using a personal laptop; Using WhatsApp to share work documents
		Justification for workarounds	Belief about limited organisational monitoring; Bypassing company restrictions (site blocking); Complexity of organisational security measures; Convenience; Heightened security measures in remote work vs office; Usability issues; Lack of attention; Generic policy rules; Privacy invasion concerns; Productivity reasons; Slow or relaxed IT response; Time constraints and task urgency; Underestimating workaround vulnerability; Underestimation of workers' capability; Cost-benefit analysis

Figure 5: Codebook of Themes and Codes.

Who is the IT Department Anyway: An Evaluative Case Study of Shadow IT Mindsets Among Corporate Employees

Jan-Philip van Acken[1], Floris Jansen[1], Slinger Jansen[1,2], and Katsiaryna Labunets[1]

[1]Utrecht University, the Netherlands
[2]LUT University, Finland

Abstract

This study aimed to explore the factors influencing employees to deploy what can be classified as shadow IT in a corporate context. Shadow IT denotes unofficial, unsanctioned forms of IT. We employed a mixed-methods approach, consisting of a survey and follow-up interviews with employees from a large professional services company. The survey yielded 450 responses, uncovering different types of shadow IT within the company. The follow-up interviews with 32 employees aimed to uncover their perceptions of shadow IT, related risks, and their attitudes towards shadow IT usage. The survey and interviews revealed various types of shadow IT and showed a dichotomy of risk-averse and risk-tolerant mindsets. We found that participants employed a combination of these mindsets. Despite being aware of significant risks, gaps exist in acting upon this awareness, leading to an awareness-action gap. Closing this gap can be facilitated through factors that change these mindsets, such as the consequences of previous shadow IT choices, risk discussions, or training.

1 Introduction

Shadow IT occurs when employees bypass official channels *"to get the IT services they want on their own"* [39]. It appears in the form of hardware, software, or services that are *"built, introduced, and/or used for the job without explicit approval or even knowledge of the organisation"* [31]. This confronts any organisation with the challenge of managing a potentially *unknown* threat introduced by well-meaning employees. Note insights from a 2021 Forbes survey, where 46% of the execu-

tives surveyed reported that "shadow IT makes it impossible to protect all of their data, systems, and applications all the time" [23]. The questions we strive to answer here are:

RQ1: *How does shadow IT usage differ between departments and ranks?*
RQ2: *What is the employee's perception of shadow IT and risks associated with its usage?*
RQ3: *Which mindset motivates employees to opt for (or against) shadow IT usage in an organisational context?*

In this paper, we conducted a mixed-method evaluative case study in one of the largest professional services organisations' branches in the Netherlands, which reports 5000+ employees, to respond to the research questions. We find that shadow IT is intertwined within the organisation's IT landscape; all types of shadow IT appeared across departments and ranks. We elicited a total of 10 risk-related mindsets that influence the shadow IT behavior of employees.

Summary of contributions: Our main contributions are:

- We present the first mixed-method study of shadow IT usage patterns, perceived implications, and specific mindsets influencing shadow IT usage.
- We quantitatively analyse the scale of shadow IT usage across different departments and ranks in a large corporate organisation through a large-scale survey of 450 employees.
- We identify four risk-averse and six risk-taking mindsets through interviews with 32 employees; combinations of these mindsets might influence shadow IT usage decisions of employees within an organisational context.
- We outline actionable recommendations for security practitioners, to improve cyber risk management in light of shadow IT based on our findings.
- By offering our aggregated survey results, interview transcripts, and codebooks as open data to the research community, we lay the foundation for future studies on shadow IT and related mindsets.

Unapproved cloud services	Use of Internet-based Software and Software as a Service (SaaS) that are not approved or unknown by the IT department. Also known as Mobile Shadow IT, once they can be accessed outside the workplace.
Self-made solutions	Use of solutions developed by employees on the company's computers to perform their work tasks. (Excel spreadsheet, application developed by employees, . . .)
Self-installed applications	Use of software installed by employees on the company's computers to perform their work tasks. (Download & installation of free of charge software from the internet, . . .)
Self-acquired devices	Use of devices owned by employees, purchased from retail rather than ordered through the official catalogue of the IT department. It includes the use of applications in the employee's personal devices at the workplace. (cf. BYOD)

Table 1: Shadow IT topology (cf. Mallmann et al. [48])

2 Background and Related work

2.1 Shadow IT Background

Shadow IT was initially viewed as an extension or support to existing IT systems [42, 73], but the misuse of official IT received attention as well [41]. Past systematic literature studies found various definitions [30, 37, 41, 42, 48]. All but one of them at least mention the definition we gave earlier by Haag & Eckhardt: *"Shadow IT is hardware, software, or services built, introduced, and/or used for the job without explicit approval or even knowledge of the organisation"* [31].

According to this definition, an employee using non-company cloud storage solutions because the client wanted the files transferred that way, or an employee building a website to help with client projects that are not officially company-supported, would thus use shadow IT.

We employed the shadow IT topology by Mallmann et al. [48], which suggested a division into four distinguished types of shadow IT to further differentiate (see Table 1).

2.2 Usable Security and Mental Models

Organisations need the ability to detect shadow IT/security and its causes. Managing cybersecurity risks should be guided by involving users rather than deploying standard solutions. Kirlappos et al. [40] and Brandon et al. [13] define actual security as *"the security provided by a system in practice, determined by (1) the security of the underlying technologies and (2) the extent to which users adopt the intended secure behaviour,"* but note that clear directives are missing.

Mental models shape behavior in specific situations. To synthesize some key findings from [66], a 'mental model' was defined as some functional internal construct that operates similarly to the process it represents [17]. According to Johnson-Laird [36], it constitutes a framework based on life experiences, perceptions, and understanding of the world. **Mental Models of Security:** Given the difficulty of defending against unknown threats, it is presumably crucial for users to stay informed about potential vulnerabilities, thus reducing

the probability of a threat actor exploiting them. Researchers in the fields of usable security and human-computer interaction increasingly rely on users' mental models to comprehend user reasoning and engagement with complex security technology. We identify two main categories in the related work.

Mental models of general security and privacy knowledge have been studied in multiple works [6, 7, 14, 49, 51, 53, 67, 69]. They highlighted that a difference in knowledge/mental models between laypeople and experts makes communication between the two groups inefficient [7]. Models are linked to metaphors and heuristics to explain said differences. While they can be useful shorthand approaches, this also explains the shortcomings: depending on the underlying simplifying heuristic, different aspects of the more complex real-world scenario are left out, in turn leaving different gaps [15]. Wash [67] showed potential dangers of security threats abusing such gaps but pointed out that "even *wrong* mental models produce good security decisions" [68]. Similar to the laypeople/experts differences, we expect different branches in the organisation to hold different mental models. Blythe & Camp [12] postulated an implementation approach for security mental models that aimed to allow for predictions of user behaviour; a valuable stepping stone when developing end-user-focused training. However, one would not need to enforce *correct* models if the mental models already present lead to *usable* security models [68].

Besides the mental models on general security aspects, multiple research works explored **mental models of specific security concepts and technologies** [1–3, 11, 19, 20, 27, 43, 46, 60, 63, 72]. They range from a study of VPN usage habits and preferences among students and general VPN users [20], over exploring adversarial machine learning mental models among practitioners [11], to examinations of the mental models of German office workers' privacy perceptions [63]. Similar to [20], we adopted a mixed-method approach to explore the shadow IT usage patterns in a quantitative survey and interviewed a diverse set of employees to gain insights into their shadow IT perceptions and related mindsets. In line with [63], we targeted the corporate population because the shadow IT phenomenon is specific to the organisational context.

Differences in Mental Models: A study by Staggers and Norcio [61] illustrated that there are big differences in the mental models of experts and non-experts, confirmed by multiple later studies [7, 25, 45]. Moreover, [7] illustrated that there is a link between the mental models of security risks and expertise in security. All these publications reported discrepancies between the mental models of participants in different groups. While previous studies indicated that non-expert would tend to be more careless and ignore warnings [14, 24, 38, 71], more recent studies showed that expertise also lead to ignored warnings, but for different reasons [54], turning security effort and potential harm into a cost-benefit consideration [5].

We explore how groups of practitioners regard cybersecurity concepts, implying differences in mindsets depending on

group composition. Understanding how to influence behavior based on a groups prevalent mindsets may facilitate protecting all end-users.

3 Research methods

To answer our research questions, we conducted a mixed-method case study. We conducted an exploratory survey to gain quantitative insights into the current shadow IT situation in the organisation (RQ1), followed by semi-structured interviews to get a qualitative understanding of survey results (RQ1) and employees' in-depth experience with shadow IT (RQ2 and RQ3). White [70] recently advocated for this integration. The benefit of doing both was that we could potentially show *if* any shadow IT was present through the survey and then follow up with an interview to assess *why* this was potentially the case and how the participants thought about the matter. Due to the anonymous nature of the survey, we could not directly link a participant's survey responses to their interview. The survey was conducted in English; interviews were in Dutch or English, depending on participant preferences.

To differentiate shadow IT types, we relied on the topology by Mallmann et al. [47] (cf. Table 1). In consultation with cybersecurity experts at the organisation, we created the following scenarios where shadow IT might occur:

S1: *(Shadow IT occurring in)* Specific client projects;
S2: *(Shadow IT occurring in)* General work tasks, so not for specific projects;
S3: *(Shadow IT occurring in)* Personal use.

The taxonomy and scenarios serve as the backbone of the survey and guide the structure and content of the questions.

Case Organisation Context: At the organisation where we conducted our study, most employees are academically trained (cf. Table 2) and work on client projects or long-term deployments, supported by the back office. Data security is crucial since the consultancy tasks they perform touch sensitive data daily.

Despite a well-defined information security policy for responsible software and hardware use (covering shadow IT management), employees enjoy some freedom to use solutions beyond the organisation's default list. These default tools are promoted through the acceptable use policy and available via the organisation's *app store*. Additional software can be downloaded, but installation requires justification through a prompt. Software usage is regularly compared against a blocklist and violations result in email notifications; we lack information on subsequent escalation steps.

Both laptops and mobile phones are managed; phones without the company portal installed on them (i.e., without an endpoint management tool) are denied access. Employees receive *cybersecurity training* during onboarding, which covers the aforementioned acceptable use policy; it states that work tasks should be performed using tools provided by the organisation. They also receive training on phishing, including

campaigns targeted at spotting and reporting attempts.

The second author was an intern at the organisation during the project, but he conducted this research independently. Besides providing input to the research team, the organisation's employees did not significantly bias the study's design or implementation, and the research team ensured the scientific rigour of the project. The organisational affiliation and internship status of the researcher were primarily logistical. They did not affect the integrity of the research process. The remaining authors have no ties to the organisation.

3.1 Survey

Design & Implementation: The core of the survey contained four main sections: the first three sections each contained three questions specific to types of shadow IT in specific scenarios (cf. Section 3). The types were *Cloud services*, *Self-installed applications*, and *Self-made solutions*, which fall outside the scope of their organisation. The fourth section covers the *Self-acquired devices* and the use of *personal emails* as types of shadow IT. These sections were refined through pilot testing, with an *Other* option provided for unlisted application types. The survey's final format emerged from an iterative feedback process. We performed three pilot rounds, each involving two new participants. Following this, we conducted a split test with a within-subjects design, comparing the two most promising survey versions. For an anonymised copy of the final survey questionnaire, see dataset. The survey was administered through the Qualtrics platform of Utrecht University[1].

In addition to the core questions, the survey included supplementary sections on informed consent, demographics and background (cf. Table 2), detailed survey instructions, and the opportunity for participants to leave their email addresses to enter a raffle and opt-in for follow-up interviews.

Recruitment: Data collection took place in April and May 2023, targeting 2000 potential participants via mailing lists, which included detailed study information, a survey link, and a flyer with a QR code. Additional printed flyers were placed in the organisation's offices, supported by direct explanations of the study's aim. Participants could win one of four prizes (two gift cards and two goodie bags), with winners announced in early June 2023.

Out of 638 initial responses, 458 were complete. To ensure data reliability, 8 outliers in completion time were excluded because their time fell beyond $\mu \pm (2\sigma)$ [35]. Overall, 70% of respondents fully completed the survey.

Participant Demographics: Table 2 provides a summary of demographics and background. Our survey's gender distribution mirrors that reported in the organisation's annual report. The hierarchical rank structure in our data reflects the pyramid shape seen in similar organisations; more lower ranks

---

[1]Utrecht University Qualtrics portal: https://survey.uu.nl

and fewer higher ranks. The spread across departments aligns with the organisation's internal distribution (cf. Table 3).

Data Analysis: Our survey included multiple-choice questions with the option for multiple answers. To analyse patterns of shadow IT usage across departments and ranks, we applied χ^2 if we had 80% of cells with values ≥ 5. Otherwise, we used Fisher's exact test, requiring the tested variables to be mutually exclusive. For this, we added a value indicating the absence of a particular shadow IT type in a group, ensuring exclusivity with instances where shadow IT was reported. We then conducted the statistical tests for each shadow IT type, answer option, scenario, and group individually. We adopted 5% as a threshold for α (i.e., the probability of committing a Type-I error). To report the effect size of observed trends, we used ϕ value, categorising the effect as *negligible* for $|d| < 0.2$, *small* for $0.2 \leq |d| < 0.5$, *moderate* for $0.5 \leq |d| < 0.8$, and *strong* for $|d| \geq 0.8$ [22]. To identify specific groups contributing to significant differences, we conducted post-hoc analyses using *residuals* for the χ^2 test [58] or *pairwise comparison* with the Bonferroni correction for Fisher's test results [59].

3.2 Interview

Interview Protocol: We followed the recommendations by [64] to create our interview protocol. The interview questions covered: (i) understanding of shadow IT, (ii) reasons for using shadow IT, (iii) perception of shadow IT usage implications, (iv) awareness of relevant organisational policies, (v) how shadow IT is discussed amongst colleagues, and (vi) how well-informed the participant feels about shadow IT.

Following Castillo-Montoya's guidelines [16], we designed our interview questions to align with our research goals. Topics were introduced before asking the main questions, and specific probes were prepared to elicit in-depth discussions on the perceived risks and implications of shadow IT. We conducted six pilot interviews with practitioners to refine our interview protocol, ensuring clarity and preventing misinterpretation. These led to only minor adjustments in question sequencing and phrasing, allowing us to include them in our final data set. For the interview guide cf. Appendix Section B.

Interviews were conducted in participants' native languages – predominantly Dutch, with two in English. Interviews were held in person; alternatively, we used Microsoft Teams, chosen for its sector popularity and its support for privacy-compliant recording.

Participants Recruitment and Demographics: We invited interview participants through an opt-in question in our survey, conducting the recruitment in two phases. Initially, we employed a dual sampling strategy for a balanced sample. *Cluster sampling* grouped participants by department (cf. department row in Table 2), while *stratified sampling* within these clusters aimed to include all ranks (cf. ranks row in Table 2). This led to 15 interviews across four departments, covering all ranks. Following an analysis of this first phase,

we sought additional participants to address gaps, such as the absence of the IT department. In the second phase, we managed to include two more from management and 15 from client-facing roles. Table 4 presents the distribution.

The 32 interviews each lasted 20-35 minutes. Limited by time and resources, we engaged with only four members of the management staff and were unable to recruit participants from the IT department. See Table 5 for participant demographics and background information.

Codebook Creation and Analysis: We used Atlas.ti [2] for open and axial coding to explore shadow IT's facets, using the interview guide to define codes and link quotes to concepts. To ensure the reliability of the results, we followed Barbour's multiple-coding approach [8], refining the codebook over six initial interviews until achieving consistent agreement between the two researchers (Krippendorff's alpha > 0.9). With this codebook, one researcher coded the remaining interviews, and the results were validated by the second researcher. All conflicts were discussed and resolved. Discussions with a third researcher in cases where the prior two could not come to an agreement ensured a correct frame of reference and minimized potential bias [28].

3.3 Ethical considerations

The Ethics Review Board of the authors' institution approved the study protocol and data management plan under reference Bèta S-23055. Participants were informed about the study details, risks, and our use of collected information before obtaining their consent (cf. Appendix Section B). Access to the survey platform was restricted to the research team and was set to exclude personal identifiers like IP addresses. Due to the sensitivity of raw survey data, only aggregated results will be published in agreement with the organisation. Interview transcripts were anonymised, participants reviewed these before final consent for publication was obtained. All personal data and raw sources were deleted post-study. Participants are referred to by numerical codes (e.g., "P03") with quotes used only from those who provided explicit consent.

4 Results & Discussion

For qualitative codes, we provide illustrative statements systematically representing corresponding themes identified across multiple interviews. These statements provide grounding for each code across all groups of participants. Section A in the Appendix provides detailed results from our qualitative analysis of the interviews. To answer RQ1, we combined the quantitative findings from the survey with qualitative insights relevant to this question identified in the interviews. The interview results covered RQ2 and RQ3.

---

[2]https://atlasti.com (23.2.1)

Gender		Age	32.7 ±9.4	Work experience	8.74 ±9.0	Rank		Education	
Male	56%	[18-25]	22%	≤5 years	45%	junior	39%	University education (WO)	86%
Female	43%	[26-30]	36%	6-10 years	24%	senior	24%	Higher Professional	
N.A.	1%	[31-40]	22%	11-20 years	15%	manager	15%	Education (HBO)	10%
		[41-50]	13%	21-30 years	11%	senior manager	12%	PhD	1%
		[50+]	7%	30+ years	5%	management	9%	Other	3%

Table 2: Summary demographics of survey participants, n=450

	Survey	Organisation*	Δ
Client-facing	84.7%	78.7%	+6.0%
Support	7.8%	16.6%	-8.8%
Management	5.8%	4.2%	+1.6%
IT	1.8%	0.6%	+1.2%

Note: a minor random noise has been introduced to the numbers in the "Organisation" column to prevent guessing the organisation's identity.

Table 3: Department Distribution Comparison

	Jun.	*Sen.*	*Mngr*	*Sen. mngr*	*Mngmnt**	**Total**
Client-fac.	6	4	5	6	-	21
Support	1	2	3	1	-	7
Mngmnt	-	-	-	-	4	4
IT	0	0	0	0	-	0
Total	7	6	8	7	4	**32**

* according to the organisational structure, all employees in the management (Mngmnt) department also holds management rank and no management employees work in other departments.

Table 4: Participant Cohort Matrix for the Interviews

4.1 RQ1: Shadow IT usage

Employees often use unauthorised external tools for work and personal tasks, with limited awareness of organisational policies. This highlights the need for cybersecurity education and communication. Self-installed applications and cloud services are extensively used, driven by the need for specific functionalities, ease of use, and overcoming IT limitations, especially in client projects.

Survey Results: Usage by Shadow IT Types Figure 1 shows the self-reported usage rate of the software-related shadow IT types across scenarios and the overall personal device usage rate. We could expect the corporate sector to be doing better; however, our survey showed a high level of shadow IT presence (up to 63%). Similarly, Gomez et al. [52] revealed a high level of shadow IT usage in US higher education in a survey of IT professionals.

Self-installed applications have a significant role in the project workflow across departments and ranks. We discovered a statistically significant use of remote workspaces ($Fisher's\ p(Fp) = 0.035$ with a small effect size (ES) ($\phi = 0.34$) for departments), conferencing tools ($\chi^2 p = 0.0025$ with a small ES ($\phi = 0.30$) for departments and $\chi^2 p = 0.00024$ with a small ES ($\phi = 0.37$) for ranks), screen capture ($Fp = 0.037$ with a small ES ($\phi = 0.35$) for departments), and

ID	Rank	Department	Degree	Age	Experience
P1	Junior	Client-facing	MSc (University)	18-25	0-3
P2	Senior	Client-facing	Postmaster	26-30	3-6
P3	Junior	Client-facing	MSc (University)	18-25	0-3
P4	Junior	Support	MSc (University)	18-25	0-3
P5	Manager	Support	Applied MSc (HBO)	51-59	30+
P6	Manager	Support	*MBO**	51-59	26-30
P7	Management	Management	MSc (University)	51-59	26-30
P8	Manager	Support	MSc (University)	36-40	16-20
P9	Senior Manager	Client-facing	Postmaster	41-50	16-20
P10	Manager	Client-facing	MSc (University)	26-30	3-6
P11	Senior Manager	Support	Applied MSc (HBO)	41-50	21-25
P12	Senior	Client-facing	MSc (University)	26-30	3-6
P13	Senior	Support	MSc (University)	18-25	0-3
P14	Management	Management	Postmaster	51-59	30+
P15	Senior	Support	PhD	41-50	16-20
P16	Manager	Client-facing	MSc (University)	31-35	7-10
P17	Junior	Client-facing	MSc (University)	26-30	3-6
P18	Junior	Client-facing	MSc (University)	18-25	0-3
P19	Senior	Client-facing	MSc (University)	26-30	0-3
P20	Senior	Client-facing	MSc (University)	31-35	3-6
P21	Senior Manager	Client-facing	MSc (University)	51-59	30+
P22	Manager	Client-facing	MSc (University)	31-35	7-10
P23	Manager	Client-facing	Applied MSc (HBO)	41-50	21-25
P24	Manager	Client-facing	MSc (University)	41-50	16-20
P25	Senior Manager	Client-facing	MSc (University)	36-40	11-15
P26	Junior	Client-facing	MSc (University)	26-30	0-3
P27	Junior	Client-facing	MSc (University)	26-30	0-3
P28	Senior Manager	Client-facing	Postmaster	41-50	11-15
P29	Senior Manager	Client-facing	BSc (University)	60+	30+
P30	Management	Management	Postmaster	51-59	16-20
P31	Management	Management	Postmaster	41-50	26-30
P32	Senior Manager	Client-facing	Applied MSc (HBO)	51-59	26-30

* - MBO stands for Secondary Vocational Education in the Netherlands

Table 5: Interview participant demographics

the "other" tools ($\chi^2 p = 0.0032$ with a strong ES ($\phi = 0.82$) for ranks) in *client-specific projects*.

Further analysis using residuals revealed that the support department uses statistically fewer **conferencing tools** (*true residual* $= -2.03$) compared to the other departments, while the management department used statistically more conferencing tools (*true residual* $= 2.02$). When looking at the nature of their work, this makes sense. *Support*, focused on internal tasks, does not need external conferencing tools beyond what the organisation provides. In contrast, the *management* is involved in landing new projects and often requires various conferencing tools like WebEx, Zoom, or Skype.

A similar test for ranks shows the lack of conferencing tools usage amongst the *junior* group (*true residual* $= -2.63$) and the extra presence amongst the *senior manager* group (*true residual* $= 2.13$). Junior employees tend to handle more hands-on work, while the latter are more involved in managing
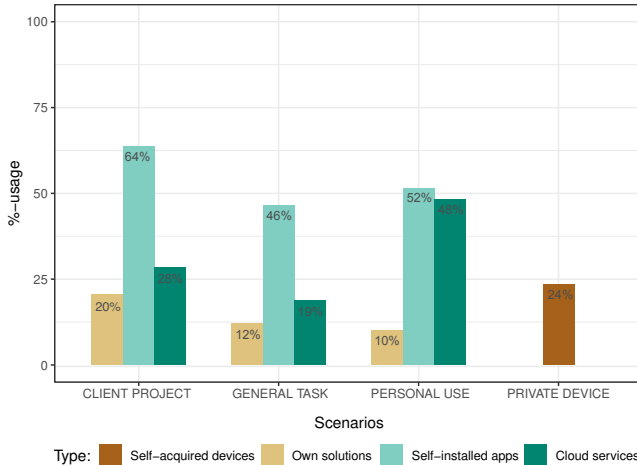
Figure 1: Rate of participants using at least one form of shadow IT. Grouped by scenarios, plus the rate of reported private device usage overall. (n=450)

projects and more frequent communication with clients.

For the *"other"* category, the post hoc test returned a residual value = 2.56, indicating statistically significant use of this category by managers. The reported examples for this category can be further categorised as: (i) data analysis tools (e.g., Azure Data Studio and R Studio) and (ii) networking and remote access tools (e.g., FileZilla, PuTTY, WinSCP, and Wireshark). The post hoc analysis did not confirm statistical significance for the rest of the types. At large, the client-facing department is mainly involved in client projects, demonstrating the biggest variability in shadow IT types used.

In general working tasks, conferencing tools ($\chi^2 p = 0.0033$ with a small ES ($\phi = 0.44$) for ranks) and streaming services ($F p = 0.0017$ with a moderate ES ($\phi = 0.70$) for ranks) stand out. For personal tasks, streaming services demonstrated statistically significant results ($\chi^2 p = 0.0072$ with a small ES (with $\phi = 0.33$) for ranks).

We find an apparent decrease in the use of self-installed applications in general work tasks vs. client-specific projects and an increase in the personal use of self-installed applications. The difference in usage of work-related and personal applications gives an initial idea of how employees see the use and hence place the potential risks of different applications on a work device, thus mitigating risks and preventing occurrences like the QQ Browser in the *management* group.

Cloud services were mostly reported for personal use. Among participants, 38-55% of the responses reported using some cloud services. *"Cloud storage"* was well represented throughout both departments and ranks. We find very high occurrences of Google Drive, Dropbox, WeTransfer, and OneDrive. This might explain why the external cloud storage services are very high in the first two work-related scenarios. If employees are used to storing and sharing files in a certain solution, they might be prone to use these in a work setting, even though

the organisation has well-supported cloud storage services. In addition to the larger cloud service providers, we see a few specialised cloud storage applications in the IT staff, like NAS solutions, with several extensions to manage and support this. Initial statistically significant results for browser extensions ($F p = 0.022$ with a small ES ($\phi = 0.35$) for departments) and browser tools ($F p = 0.053$ with a moderate ES ($\phi = 0.66$) for ranks) were not confirmed by the post hoc analysis.

Self-made solutions: Employees find the need to *create their own solutions* sometimes, indicating a gap between their unique needs and the tools provided by their organisations. They demonstrate resourcefulness and creativity in using their own software, websites, external spreadsheets, and system couplings, among other solutions. Solutions span from niche calculations to tracking spreadsheets, forecasting models, and task automation. Across all roles, we find that self-built solutions are lower in personal contexts, suggesting that they are driven by work-related needs rather than personal preferences. The patterns imply that all employees, regardless of role, encounter tasks for which existing systems do not offer standard solutions. However, the statistical tests did not reveal any significant differences across cohorts.

Private devices/emails: We do not observe a lot of usage of private devices or personal emails (76% of participants reported no usage), and it is well spread across different ranks and departments. Among users, we identified two prevalent cases: using private devices/emails for 'mailing and communication', including emailing colleagues, forwarding emails to personal accounts, and calling clients and candidates, and 'calendars and reminders', where work and private calendars are sometimes merged, and private reminders can be set for work. Statistical tests did not reveal any significant differences across cohorts.

Interview Results To complement the survey findings, we conducted semi-structured interviews and analysed the reasons behind shadow IT usage and usage-related patterns. We now discuss our qualitative findings.

Reasons for using shadow IT: The main reason, reported by 10 out of 32 (10/32) participants, was the *need for specific functionality* since a participant seeks a certain functionality not covered by the approved solutions: *"Well, it is often for work-related matters that there is no such thing within the current tools [. . .]"* [P3] Another top reason is *client requirements* (4/32) when the participant had to use a certain shadow IT application because of a client project. On installing unofficial programs, the system prompts for a justification, aiming for conscious decisions as to why users install unsupported applications: *"Yes, I have always used them for client projects. I have never installed anything that I did not need for a client project [. . .] Whenever we have to install something from an unknown source, the system wants you to enter a reason why you are installing this application. For me, the reason is*

always to support a client project" [P22]

The other reasons are related to the employees' *habits* (8/32) when they work with a certain software for years: *"[...]I have worked with it for years, so then it also becomes a habit, and I'm happy with it"* [P5] or because *these tools allow me to do it quickly and easily."* [P3]

A *workaround* (5/32) as a means to get some tasks finished is also mentioned: *"[...] So we just want the functionality, just the tool. If a website is blocked, but you need to access it, or you do want to send that email, you grab your phone, where it is not blocked, or you use another device or browser. If they really need it, people will find a way"* [P24]

Among the less frequent reasons, we found *insufficient standard solutions*, *time constraints*, *financial feasibility*, and overcoming a *language barrier*. Our findings are aligned with [50] and [18], who emphasised the occurrence of shadow IT to address deficiencies in official IT systems and provide additional reasons for these occurrences.

Policy/awareness/usage gap: Despite the organisational policy forbidding external tools for business-related tasks, Figure 1 shows that employees use them a lot for both work and personal tasks. Implicitly, the policy allows using external tools for private tasks, placing trust in the employee adherence. Monitoring can only detect the tools' usage *in general*, unable to tell private or business tasks apart. These services are used across all job levels and situations, aggregated by ranks and departments. We asked our interviewees what shadow IT means to them and how aware they are of the related organisational policy. Only six participants (6/32), all from the *client-facing department*, were able to define what shadow IT is, while 15 participants demonstrated familiarity with the related policy: *"You may only use applications that are approved by [organisation]. I mean they have the [internal app store] for a reason, right, in addition to a whole protected environment with work applications and services."* [P2]

Participants generally felt informed about shadow IT implications, yet their actions sometimes contradicted this knowledge, highlighting an *awareness-action gap*. Hielscher and Parkin [34] found that effective security awareness programs are often constrained by a lack of clear goals and communication between managers and employees. Shadow IT usage decisions involved evaluating both internal (company-provided) and external (shadow IT) solutions, often requiring consultation with IT teams and higher-ranked individuals, reflecting the organisation's hierarchical structure.

Among the policy-aware participants, we saw that even though there are ways to turn shadow IT into *accredited* IT[3], those seem to be taken only rarely: *"I don't even know who IT is, and with that comes the risk that you might receive a 'no' to your request. Meaning you cannot do the engagement,*

while needing the functionality. So by approaching IT, you enter a negotiation you need to win [...]" [P22]

A clear split emerges among those *unfamiliar* with the policy (17/32). Some openly admit they do not know it, while others list their cybersecurity courses to prove their knowledge. Some talk about different policies or show other proof of their cybersecurity know-how. It appears cybersecurity is perceived to be crucial in the workplace, and not being up to speed can lead to significant consequences: *"Ehm, I think I should be familiar with this. I think it says something like just use your common sense when handling technology, right?"* [P13] One participant had a moment of realisation that the provided information in the interviews could be self-incriminating, in the sense that participants provide information with regards to not being aware of cybersecurity standards: *"Ah, now I understand why this interview is anonymous. I think you will really get punished for this kind of stuff"* [P25]

Most of our participants believed that they are reasonably (14/32) or well-informed (7/32) about the use of technology: *"We do have the mandatory courses which teach us all sorts of things that can go wrong. So we need to be very aware. I think there is a great awareness of how to handle things in this regard"* [P5] Some demonstrated adequate knowledge of how to act but did not always follow through: *"I have a good idea of what I should and should not do. However, I do not always fully act like it [...]"* [P16]

Shadow IT Perspective: The *"perspective"* of shadow IT refers to a viewpoint shift when working on a client project using client-provided resources: *"So that would mean we would need an environment at the client. This can be a client laptop or environment through [remote workspace]. Both with the tools installed so that the client pays for licenses and puts the responsibility for updating on the client. Moreso to put the risk of these applications in their shoes."* [P22]

This shift places all applications on the client's system outside the organisation's purview. It is commonly observed among employees engaged in client-facing roles, with examples including the use of client laptops, remote workspaces, and client licenses to create distinct work environments.

We observe that participants doing longer projects for a single client often get a physical device in the form of a *client laptop* (7/21)[4]. They need to do a mini-onboarding process to install all relevant software, but in doing so, they mitigate any shadow IT threats for their own organisation:

"Sometimes we work on laptops provided by the client. And then the rulebook changes because it is theirs, so then you have a lot of contact with the client's IT team." [P17]

In other cases, the employees might get access to a *remote workspace* (4/21) or the client shares the *licenses* for the necessary applications (3/21): *"The client uses a certain ap-*

---

[3]Present the application to an online service desk, once it passes checks (licenses used, vendors, compliance, etc.), it counts as accredited.

[4]Here we report code grounding within the client-facing group as this observation is specific to this group.

plication, we will copy that and just work from their accounts in those systems." [P10]

4.2 RQ2: Shadow IT Perception

> We observed a complex interplay between perceived risks, benefits, and mitigation strategies in shadow IT usage. Shadow IT is used for efficiency gains and cost savings despite awareness of cybersecurity risks like malware and data leaks. Usage strategies seem influenced by participants' mindsets.

Perceived Benefits: Across cohorts, we find a nuanced perception of shadow IT. Participants discussed perceived benefits, noting that certain tools allow them to work more *efficiently* (8/32): *"[Tool] can be used for a variety of things. For example, [. . .] I had to do [working task], I then used that tool, and it just saves me so much time"* [P27] This observation aligns with Pinto et al. [18], who demonstrated that both workaround behavior and shadow IT usage positively impact individual performance.

Participants also consider financial feasibility (*cost benefit*; 3/32) as a reason to use shadow IT: *"[. . .] the costs for the organisation. I mean, if everyone went to IT for every small thing, that would not work [. . .]"* [P16]

Perceived Risks: Our participants frequently related *data leaks* (23/32) to the use of shadow IT: *"My biggest concern is and always will be data breaches. So this is when we consciously send our data somewhere we cannot oversee the risks anymore"* [P30]. Sometimes, they described the concept rather than explicitly naming it: *"In general, it is quite hard to find out who is behind the tool and what exactly they do to your data [. . .]"* [P25]

As a precursor to data leaks, the participants often mentioned the *malware* threat (14/32). The only specific type of malware that was mentioned is a virus. However, some participants described infected or malicious programs:

"I suppose it could lead to viruses [. . .] This can then lead to the access of certain data on your laptop" [P23]

"When you download software that could just be malware, this can infiltrate your computer. This opens up the [organisation] network, and then anything can happen to data" [P3]

Next to malware, *unauthorised access* (10/32) was perceived as possible risk due to shadow IT. Some participants explicitly state a threat actor, while others simply focus on unauthorised access by *some entity*:

"The most important danger is giving access to others. Access that allows them to access data that they shouldn't" [P10]

"Hackers can get access to our system, and then they can access sensitive data from clients [and] exploit this data." [P15]

Non-central governance (8/32) concerns the principle at the core of the potential threats related to shadow IT, even preceding the malware and unauthorised access. Namely, shadow IT instances fall outside of the scope of the organisation, and

therefore, the organisation can not perform standardised cybersecurity checks on these instances:

"What if you were to download something that is monitored by your employer, you could always get an alert or notification that says, hey, something is wrong here. So if you go outside of the employer, you bypass all checks and expose yourself to vulnerabilities" [P1]

"The disadvantage is always that if it is not checked by [organisation], even if there might be very evident risks, they will not be aware of this" [P19]

As expected, managers prioritise the organisation's reputation (5/32): *"The biggest risk of all is the reputation damage for [organisation] due to data breaches. Since all the work we do is confidential, and sometimes even holds price-sensitive information"* [P14] It is noteworthy that even junior employees show a high level of awareness about *reputation risk*, suggesting widespread awareness: *"So if we do something that makes [organisation] untrustworthy, this can impact the name and therefore everyone in the organisation"* [P4].

Among the less evident risks of shadow IT, we also encounter *ransomware* (3/32) and *misinformation* (2/32). The latter is mentioned in light of the recent rise of generative Large Language Models: *"[. . .] if you just copy the answers as they come out and if you do not use your common sense anymore, naively thinking that the answer is always true, that is a big risk"* [P11].

Contradictions: We observed several examples where participants illustrated risk comprehension, yet rationalised their own use of shadow IT instances. Specifically, we encountered two situations where participants (identifiers changed) realised that paradox and figured that their behavior was in conflict with the organisation's protocols:

"Yes, you should always be careful with these things. [. . .] We might be already crossing a line here. [. . .] Now, thinking about it this way, I do not think it is allowed. Because it is not a [organisation] tool." [P11]

"We just need to put this into practice. So perhaps I should ask my supervisors about our usage of tools outside the [organisation] toolbox" [P15]

4.3 RQ3: Mindsets of shadow IT usage

> Our study revealed a dichotomy in attitudes towards shadow IT, with four mindsets favouring risk aversion and six inclined towards risk tolerance. Individuals' approaches to shadow IT are influenced by evolving mindsets, contexts, and experiences, highlighting the complexity of decision-making in this area and the impact of external factors like discussions or awareness-raising initiatives.

When coding interview transcripts, we noted instances where participants subtly illustrated their conceptualisation of shadow IT. We identified codes representing various mind-

sets and their interplay, reflecting internal drivers influencing participants' attitudes towards shadow IT usage. We found ten distinct mindsets: four risk-averse (35 coded statements among 23 participants), promoting cautious behaviours when dealing with shadow IT, and six risk-taking (37/22), increasing individuals' risk appetite with regard to shadow IT.

4.3.1 Risk-Averse Mindsets

RA1. Consequence-Avoidance Orientation is a mindset where individuals prioritise steering clear of negative outcomes or consequences when making decisions and taking action. We found 17 participants demonstrated a high awareness of various consequences and are therefore cautious to avoid potential negative impacts. *"Think about all the consequences. I think those hold the biggest risks. Which is also the reason I don't have anything external."* [P19]
This is potentially related to prolonged exposure to a tool/service that was perceived as 'bad' by users, thereby lessening their well-being score [21], with the extreme case arguably being complete avoidance.

RA2. Knowledge-Based Conservatism mindset (8/32) is defined by a preference for using established knowledge and wisdom as a basis for decision-making. We noticed that a specific group of participants showed a notably higher level of awareness regarding the concept of shadow IT. They also demonstrated a deeper understanding of the associated risks and implications, thanks to their extensive expertise in information technologies. This equipped them with the knowledge to navigate the challenges posed by shadow IT effectively. We related this mindset to participants' expertise in technology, which influenced more secure behaviour in the context of shadow IT: *"I am very aware of all sorts of risks. It is because of my role as [role]. So, therefore, I am aware of certain things that the average Joe here won't think of"* [P7]

RA3. Risk Transfer Mindset (8/32) is characterised by a tendency to transfer risks to external/other entities. In our study, we observed participants within the *client-facing* group try and manage any shadow IT consequence by shifting the *perspective* of shadow IT to clients. This strategic approach makes it more convenient for clients and helps to mitigate potential shadow IT threats for the organisation.

"I would let the client take responsibility for the risk. Because they are the ones asking for this tool. However, I would not have thought of that when I was younger." [P22]

RA4. Cautious Seasoned Judgement mindset (4/32) reflects a thoughtful decision-making approach informed by broad experience, similar to but distinct from *Knowledge-based Conservatism*, which relies on specific expertise. This mindset is not consciously used to guide shadow IT actions, but it manifests in individuals who, similar to a 'cautious seasoned judge', encourage colleagues to appreciate the value of accumulated wisdom and experience in making well-thought-out decisions. Our observations suggest that individuals' com-

binations of mindsets, incl. learning from past shadow IT usage outcomes, can evolve over time. This mindset can be compared to the practice of introducing security and privacy champions [9, 62] who care about security and might act as those "seasoned judges".

"I have seen it all, but actually you should go through a data breach once just to see how bad it really is. After that, you'll think twice about your actions. You learn this through trial and error over the years." [P30]

4.3.2 Risk-Taking Mindsets

RT5. Common Sense Fallacy mindset, prevalent among our participants (11/32), revolves around the idea that discussions about shadow IT and cybersecurity, in general, should be minimal due to an assumed baseline of 'common sense' understanding. Those holding this view believe that individuals should already grasp fundamental cybersecurity concepts.

Individuals with this mindset intuitively know what is acceptable or not within a given context. It is crucial to recognise that not everyone possesses this basic cybersecurity knowledge. Assuming universal understanding can reduce important team discussions, a drawback when dealing with shadow IT. While 'common sense' facilitates decision-making, it can also sideline critical conversations, adversely impacting overall shadow IT behaviour: *"In our department, they just expect you to know this stuff. You need to have a certain knowledge of these things. I mean, you follow a certain education, and you get all these e-learnings."* [P18]

RT6. Illusion of Sufficiency mindset (6/32), wherein an individual erroneously believes they do not require any shadow IT applications, under the assumption that all necessary tools are already provided by their organisation. This notion is exemplified by citing instances of shadow IT, effectively illustrating the 'illusion'. Consequently, individuals with this mindset tend to perceive themselves as immune to related risks, assuming all solutions are sanctioned by their organisation. This misbelief diminishes their vigilance towards potential cybersecurity threats.

It is noteworthy that all participants holding this mindset exhibited a lack of familiarity with shadow IT. This knowledge gap perpetuates the misconception that all the tools they employ are officially endorsed, which may not be the case. This attitude characterises the essence of this mindset: *"No, for me this is not a thing to consider because we have everything taken care of."* [P6]
"[...] I think in terms of work-related things we have everything that we need." [P19]

RT7. Misguided Sense of Protection (6/32) Individuals hold a false or erroneous belief in their own protection. This mindset is noticeable in our participants, many of whom manifest insecure norms. Participants often recount their experiences with other security measures, such as those addressing phishing and viruses, and consequently, they extrapolate that these

protections extend to safeguarding them against shadow IT.

Consequently, these individuals possess a sense of invincibility, perceiving that the organisation's protection shields them from harm. In the realm of cybersecurity, this excessive perception of invulnerability influences participants' behaviour concerning shadow IT usage. They operate under the false premise that any unauthorised usage or installation would trigger alerts, creating a *false sense of security*: *"[. . .] I think they watch what you downloaded, and if it is not okay then maybe it will go through a system that detects this, or maybe there is a team that reads everything, and you then get a message to delete it from your machine"* [P15]

"And also you get a warning I think at [organisation] if you have something on your system which is not good [. . .]" [P5]
Behaviour akin to this mindset has been pointed out as potentially dangerous [46], citing that erroneous user mental models of systems *"expose users to security and privacy risks.*"

RT8. Performance-Driven Rule Bending mindset (5/32) centred on achieving specific outcomes, even at the expense of adhering to established rules and guidelines. Participants occasionally demonstrate a readiness to disregard or actively circumvent standard cybersecurity protocols to meet work deadlines. This negatively impacts the overall shadow IT behaviour of individuals:

"I cannot explain to a client that certain tasks have not been completed. This means that sometimes employees enter a grey area, perhaps even cross it by doing what they shouldn't. I think everyone is aware of this [. . .]" [P20]

"the main issue is that the show must go on [. . .]" [P20]

RT9. Longevity-Based Invincibility (5/32) Individuals believe that the extended presence of a concept grants them a sense of immunity from adverse effects. This form of survivorship bias leads participants to disregard potential negative outcomes associated with shadow IT, mainly due to their positive long-term experiences with these solutions, fostering a perception of 'invincibility.'

We have observed instances where entire teams have adopted specific shadow IT solutions for an extended period, fostering an illusion of safety among them. Consequently, new employees, introduced to these tools as a longstanding practice, may not fully grasp the associated risks. The attitude of *"we've used it for so long without any issues"* represents this mindset and erodes their vigilance in managing shadow IT instances effectively: *"[. . .] I don't know, I think sometime a while ago it was introduced, and it has stayed up until now [. . .] over time it has grown to what it is now for us."* [P12]

RT10. Cost-Driven Compromise (4/32) Individuals make decisions based on financial considerations. We have observed a clear pattern among participants, wherein cost savings are explicitly prioritised in their shadow IT decisions. The *"we use it because it is free"* attitude represents this mindset, and it significantly undermines the shadow IT behaviour of individuals: *"I wonder about, for example, [tool], since we used it because it provides a free package. One might wonder how good*

(a) Per Department

(b) Per Rank
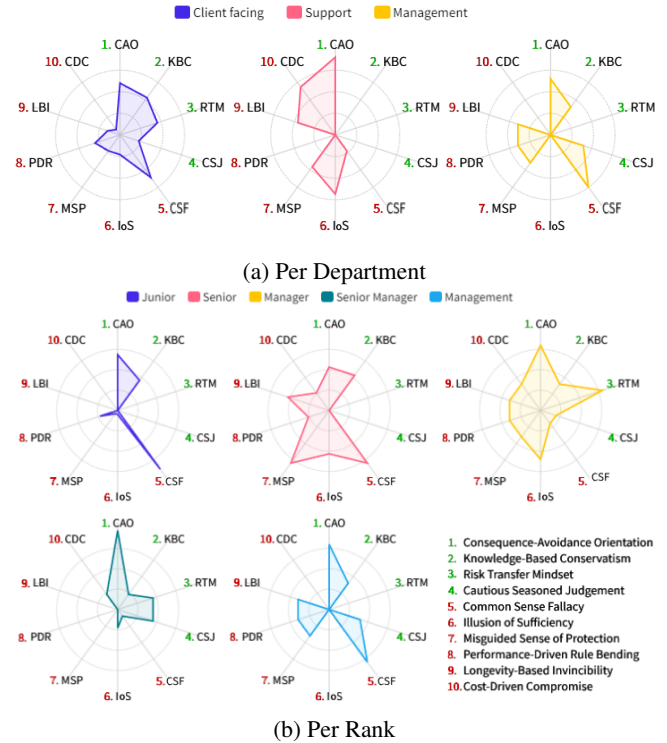
Figure 2: Relative Occurrence of Mindsets

that is [. . .]" [P5] Security mindsets and organisational security culture are shaping employee behaviour and adopted practices [32]. Schoenmakers et al. [57] revealed that the security mindset involves aspects like proactive monitoring, investigating, and evaluating potential security threats. In our study, we focused on risk-taking and risk-averse mindsets, but similarly to those aspects, our mindsets can potentially manifest at different levels and combinations in employees. Moreover, Ryan et al. [55] identified four security archetypes that are similar to our RA1 and RA3 (or "pragmatics"), RA4 (or "champions"), RT9 (or "optimist"), RT7 and RT8 (or "heroes").

4.3.3 Mindset Patterns

We explored the occurrence patterns of certain mindsets across departments and ranks. Figure 2 visualise two cohorts through radar graphs, with different mindsets per axis. The data is normalised to account for varying cohort sizes, focusing on relative occurrences (denoted by the rings in the figures) to identify patterns across different groups.

Departments: Across different departments, we observe distinct patterns of mindset presence, as illustrated in Figure 2a. While four mindsets (RA1, RT5, RT7, and RT9) are prevalent across all departments, most are of a risk-taking nature, suggesting a lack of a general organisation-wide risk-averse mindset. Notably, the *client-facing* and *management* groups exhibit similarities in the combination of risk-averse (RA1,

RA2, RA3) and risk-taking mindsets (RT5, RT7, RT8) due to their overlapping work responsibilities. A mild distinction arises when comparing these two groups with the *support* department, mainly attributable to the absence of risk-averse mental models (except RA1) in the latter.

Ranks: Regarding employee ranks, we identify a few light trends (see Figure 2b). Notably, the risk-averse mindsets RA1 and RA2 are consistently present across all ranks, indicating a widespread awareness of potential shadow IT consequences and expertise that positively influence shadow IT behaviour. Moreover, the risk-averse mindsets RA3 and RA4 were found to be prevalent among higher ranks, such as *manager*, *senior manager*, and *management*. These mindsets align well with the responsibilities and challenges these employees face, suggesting that individuals in higher-level roles adopt a more risk-conscious approach to shadow IT decision-making. This highlights the importance of involving these groups in work scenarios.

5   Implications of Findings & Limitations

In this section, we discuss the implications of the findings reported in Section 4 and provide recommendations for practitioners. Moreover, we discuss the limitations of the study.

Research Implications: Our study explored shadow IT usage and employees' perceptions and attitudes within a large corporate setting. It identified ten key mindsets affecting employees' perceptions of and decisions for shadow IT. The results reveal how these mindsets affect shadow IT behaviours, resulting in risk-averse or risk-taking behaviours in employees.

While no major patterns linked specific cohorts to particular mindsets, the variation across employee groups highlighted the need for broader research across diverse populations to capture the full spectrum of mindsets present. Our portfolio of mindsets can inform future qualitative and quantitative research among various populations and contexts, serving as a foundational framework.

Recommendations for Practitioners: To address the challenges related to shadow IT, based on our findings, we suggest:

- **Transparent Communication:** By fostering an environment where employees feel safe and comfortable discussing their technology needs, organisations can identify and address potential shadow IT instances. Not only can this approach mitigate usage risks of shadow IT, it also builds trust between the IT department and other employees, creating a more cooperative and secure digital work environment.
- **Targeted Shadow IT Awareness Training:** The interviewees often related the knowledge of potential negative consequences of shadow IT to recurring mandatory training. While we did not quantify this trend, the consistent focus on threats and repercussions positively affected employee behaviour, particularly reflected in *Consequence-Avoidance Orientation* and *Knowledge-Based Conser-*

vatism mindsets (see Section 4.3). To foster awareness of shadow IT consequences, we recommend maintaining a training initiative. Previous work [33] suggests an untapped benefit here: educating users about the capabilities of a tool could increase usage, thus potentially boosting the usage of official tools. However, we found examples of the *support* group who had to do training that was not relevant to them and therefore the overall perception of training is seen as less important. Thus, we recommend tailoring training content to align with the specific needs of various roles, departments, and mindsets, ideally tapping into existing functional models or using functional metaphors, as has been suggested [56,72].

- **Shadow IT Protocols:** For the employees with *Performance-Driven Rule Bending* mindset, we suggest creating protocols supporting individuals in navigating rule-bending situations while making it as safe as possible. This approach is supported by Pinto et al. [18], who argue that while shadow IT poses certain risks, it also provides significant benefits to individual performance.
- **Track Long-term Instances:** To manage the *Longevity-Based Invincibility* mindset, the IT team could track down the use of long-term-adopted shadow IT tools and uncover their adoption reasons. From there, an informed decision about phase-out, replacement, or take-over can be taken, as per Fürstenau et al. [26].

Finally, we stress the significance of transparent communication regarding information security policies. We emphasise the need to accommodate employees' perspectives and needs in the supplied software, hardware, and training opportunities.

Future work: Given the number of interviewees and the specific organisation targeted here, replicating our research to confirm our findings is viable. We encourage further exploration into how different mindsets converge in decision-making, potentially through a controlled game-like scenario for rich data collection. Understanding how combinations of mindsets impact the resulting shadow IT behaviour would prove valuable to boost overall secure behaviour. This could theoretically be achieved by either hampering risk-taking mindsets or strengthening the risk-averse mindsets.

We assume that individuals can hold a combination of various mindsets related to shadow IT. We apply the '*theory of planned behaviour*' [4] and assume that for users to have the intention to display safe security behaviour, there are three main contributing factors. In no particular order, we have first the subjective norm regarding the behaviour, the pressure by one's surroundings to engage in or abstain from a behaviour. Secondly, there is the perceived behavioural control; and lastly, the mindset towards the behaviour (called attitude in [4]). We have yet to uncover to what extent certain mindsets are present or how these are influenced by differing situations.

We observed that shadow IT instances are not limited to individual users but also involve departments or smaller groups

within an organisation (e.g., *Longevity-Based Invincibility* mindset). While we have taken certain cohorts to horizontally and vertically divide the employee group for analysis, we have not seen obvious patterns of shadow IT instances across the chosen cohorts. Future research might uncover what different cohorts provide the most optimal division of individual groups, such that clear patterns of combinations of shadow IT mindsets in certain groups become apparent.

Limitations: To support our readers in an appropriate contextualisation of our results, we discuss the key limitations and describe how we reduced their impact. The survey's respondent composition mirrors the larger organisation, leading to a prevalence of *client-facing* group responses and fewer from other groups. This difference in group sizes may challenge the χ^2 test's validity, which requires over 80% of cells to have values above 5. Thus, we adopted Fisher's exact test for scenarios where the χ^2 test's assumptions were not met.

Our survey design might be lengthy and holds some nuances. We split the survey into four shadow IT types and set it in three scenarios. Thus, if respondents do not read the explanation and context carefully enough, the responses can be prone to errors. To prevent errors, we placed clear instructions after the demographics part, emphasising the focus on applications beyond the organisation's norm and clarifying scenario contexts, and validated these changes in a pilot test.

Another limitation of the survey was receiving challenging explanations, such as *support* department participants, who mainly deal with internal tasks, answering sections on client-specific projects—possibly indicating misunderstandings or inattention. However, given the support group's small portion (7.8%) of our sample, its impact seems minimal, aside from their notably lower use of conferencing tools (see Section 4.1).

Given the sensitivity of the shadow IT topic, our study might be prone to social desirability bias (SAB), prompting participants to provide socially acceptable rather than truthful responses. To mitigate it, the survey was anonymous [29], the interview was kept confidential, indirect questioning was used [44]. Our interviewees expressed freedom to express "undesirable" opinions without withdrawing from the interview (see P25 at Section 4.1), proving that the researcher was able to establish trusting rapport with the participants [10]. For this, we developed a uniform protocol to probe participants' shadow IT perceptions, observing varied question comprehensions among participants from different groups. To maintain the interview flow, we sometimes provided application examples, which may have influenced responses.

A significant limitation is our failure to interview IT department staff, although some were surveyed. Future research should bridge this by interviewing IT professionals to understand their shadow IT mindsets, similar to what has been done regarding 'the' security mindset [57]. While our study reflects large corporate environments, broader validation across different organisational contexts is recommended.

Our study did not examine the impact of the growing trend towards remote work on shadow IT. This evolving work dynamic calls requires further investigation to understand its effects on shadow IT practices within organisations, offering important insights for both academia and industry.

6 Conclusion

This work investigated the perception of the shadow IT concept: the occurrences of shadow IT, how its usage varies across different cohorts in a large organisation, and the mindsets associated with it. We find that shadow IT is an intertwined part of the organisation's IT environment, observing all types differentiated by [47]: *cloud services*, *self-installed applications*, *self-built solutions*, and *personal devices*. We notice that users opt for familiar tools and services to meet work or personal needs; if these tools are not provided by default in the organisation then users tend to opt for shadow IT.

Most threats associated with shadow IT are perceived differently across cohorts, reflecting varying degrees of risk awareness and differing risk-mitigating approaches. Despite this awareness, we found inconsistencies and gaps in acting upon this awareness, resulting in an *awareness-action gap*.

The understanding and perception of shadow IT across cohorts are conceptualised through ten different mindsets. We differentiate *risk-averse* mindsets from *risk-taking* mindsets and propose that individuals typically hold a combination of these based on several personal and work-related factors and their current context. We consider that a combination of these mindsets influences individual shadow IT behaviour.

This research provides comprehensive and practical insights into employee perceptions of shadow IT. It points towards shadow IT's dichotomous nature: *a push towards non-standard solutions for efficiency and cost reasons, balanced against a broad awareness of significant risks*.

To manage the challenges related to shadow IT, we recommend the following measures: (i) fostering an environment where employees can openly discuss their technology needs, (ii) maintaining high awareness through tailored training, (iii) creating shadow IT protocols for certain scenarios, and (iv) tracking long-term shadow IT instances and conducting their risk assessment. This investigation of shadow IT, while providing practical insights and recommendations, also identifies the need for future work in understanding the behavioural impact of the combination of shadow IT mindsets. By exploring these implications, organisations can better manage shadow IT, minimising potential risks while maximising the benefits.

Data Availability: Data (incl. survey and interview questionnaires, aggregated survey results, summarized demographic information, and de-identified transcripts) is made available via Utrecht University data publication platform Yoda for a minimum period of 10 years [65].

References

[1] Noura Abdi, Jose M. Such, and Kopo M. Ramokapane. More than Smart Speakers: Security and Privacy Perceptions of Smart Home Personal Assistants. In Heather Richter Lipford, editor, *Proceedings of the 15th Symposium on Usable Privacy and Security (SOUPS)*, pages 451–466, Santa Clara, CA, USA, 2019. USENIX Association.

[2] Ruba Abu-Salma, Elissa M Redmiles, Blase Ur, and Miranda Wei. Exploring User Mental Models of End-to-End Encrypted Communication Tools. In Lex Gill and Rob Jansen, editors, *Proceedings of the 8th USENIX Workshop on Free and Open Communications on the Internet (FOCI)*. USENIX Association, 2018.

[3] Ruba Abu-Salma, M. Angela Sasse, Joseph Bonneau, Anastasia Danilova, Alena Naiakshina, and Matthew Smith. Obstacles to the Adoption of Secure Communication Tools. In Úlfar Erlingsson and Bryan Parno, editors, *Proceedings of the 38th IEEE Symposium on Security & Privacy (S&P)*, pages 137–153. IEEE, 2017.

[4] Icek Ajzen. The Theory of Planned Behavior. *Organizational Behavior and Human Decision Processes*, 50(2):179–211, 1991.

[5] Devdatta Akhawe and Adrienne Porter Felt. Alice in warningland: A Large-Scale field study of browser security warning effectiveness. In *22nd USENIX Security Symposium (USENIX Security 13)*, pages 257–272, Washington, D.C., August 2013. USENIX Association.

[6] Bilal Al Sabbagh and Stewart Kowalski. Developing social metrics for security modeling the security culture of it workers individuals (case study). In *Proceedings of the 5th International Conference on Communications, Computers and Applications (MIC-CCA)*, pages 112–118. IEEE, IEEE, 2012.

[7] Farzaneh Asgharpour, Debin Liu, and L. Jean Camp. Mental models of security risks. In Sven Dietrich and Rachna Dhamija, editors, *Proceedings of the 12th International Workshop on Usable Security (USEC)*, volume 4886 of *Lecture Notes in Computer Science*, pages 367–377. Springer, 1 edition, 2007.

[8] Rosaline S Barbour. Checklists for improving rigour in qualitative research: a case of the tail wagging the dog? *British Medical Journal*, 322(7294):1115–1117, 2001.

[9] Ingolf Becker, Simon Parkin, and M Angela Sasse. Finding security champions in blends of organisational culture. *Proc. USEC*, 11:124, 2017.

[10] Nicole Bergen and Ronald Labonté. "everything is perfect, and we have no problems": detecting and limiting social desirability bias in qualitative research. *Qualitative health research*, 30(5):783–792, 2020.

[11] Lukas Bieringer, Kathrin Grosse, Michael Backes, Battista Biggio, and Katharina Krombholz. Industrial practitioners' mental models of adversarial machine learning. In Apu Kapadia Sonia Chiasson, editor, *Proceedings of the 18th Symposium on Usable Privacy and Security (SOUPS)*, pages 97–116. USENIX Association, 2022.

[12] Jim Blythe and L Jean Camp. Implementing mental models. In Lorrie Faith Cranor, editor, *Proceedings of the 8th Symposium on Usable Privacy and Security (SOUPS)*, pages 86–90. USENIX Association, 2012.

[13] Merel Brandon, Hanna Kathrin Schraffenberger, Wouter Sluis-Thiescheffer, Thea van der Geest, Daniel Ostkamp, and Bart Jacobs. Design principles for actual security. In *Proceedings of Nordic Conference on Human-Computer Interaction (NordiCHI)*. ACM, 2022.

[14] Cristian Bravo-Lillo, Lorrie Faith Cranor, Julie Downs, and Saranga Komanduri. Bridging the gap in computer security warnings: A mental model approach. *IEEE Security & Privacy*, 9(2):18–26, 2010.

[15] L Jean Camp. Mental models of privacy and security. *IEEE Technology and society magazine*, 28(3):37–46, 2009.

[16] Milagros Castillo-Montoya. Preparing for interview research: The interview protocol refinement framework. *The qualitative report*, 21(5):811–831, 2016.

[17] Kenneth James Williams Craik. *The nature of explanation*, volume 445. CUP Archive, 1967.

[18] Aline de Vargas Pinto, Iris Beerepoot, and Antônio Carlos Gastaud Maçada. Encourage autonomy to increase individual work performance: the impact of job characteristics on workaround behavior and shadow it usage. *Information Technology and Management*, 24, 2023.

[19] Constanze Dietrich, Katharina Krombholz, Kevin Borgolte, and Tobias Fiebig. Investigating system operators' perspective on security misconfigurations. In Michael Backes and XiaoFeng Wang, editors, *Proceedings of the 25th ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 1272–1289. ACM, 2018.

[20] Agnieszka Dutkowska-Zuk, Austin Hounsel, Amy Morrill, Andre Xiong, Marshini Chetty, and Nick Feamster. How and Why People Use Virtual Private Networks. In Kurt Thomas Kevin Butler, editor, *Proceedings of the 31th USENIX Security Symposium (USENIX Security)*, pages 3451–3465. USENIX Association, 2022.

[21] Sindhu Kiranmai Ernala, Moira Burke, Alex Leavitt, and Nicole B. Ellison. Mindsets matter: How beliefs about facebook moderate the association between time spent and well-being. In *Proceedings of the 2022 ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 2022.

[22] Christopher J Ferguson. *An effect size primer: A guide for clinicians and researchers.* American Psychological Association, 2016.

[23] Forbes Insights. Perception gaps in cyber resilience: Where are your blind spots? *Forbes*, 2021.

[24] Batya Friedman, David Hurley, Daniel C Howe, Edward Felten, and Helen Nissenbaum. Users' conceptions of web security: a comparative study. In Dennis Wixon, editor, *Proceedings of the 2002 ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 746–747. ACM, 2002.

[25] Steve M Furnell, Peter Bryant, and Andrew D Phippen. Assessing the security perceptions of personal internet users. *Journal of Computer Security*, 26(5):410–417, 2007.

[26] Daniel Fürstenau, Hannes Rothe, and Matthias Sandner. Leaving the Shadow: A Configurational Approach to Explain Post-Identification Outcomes of Shadow IT Systems. *BUS INF SYST ENG*, 63, 2020.

[27] Kevin Gallagher, Sameer Patil, and Nasir Memon. New Me: Understanding Expert and Non-Expert Perceptions and Usage of the Tor Anonymity Network. In Sonia Chiasson and Matthew Smith, editors, *Proceedings of the 13th Symposium on Usable Privacy and Security (SOUPS)*, Santa Clara, CA, USA, 2017. USENIX Association.

[28] Lucia Garcia and Francis Quek. Qualitative research in information systems: time to be subjective? In Janice I. DeGross Allen S. Lee, Jonathan Liebenau, editor, *Proceedings of the 8th International Federation for Information Processing (IFIP)*, pages 444–465. Chapman & Hall, Ltd., 1997.

[29] Ahmad Nauman Ghazi, Kai Petersen, Sri Sai Vijay Raj Reddy, and Harini Nekkanti. Survey research in software engineering: Problems and mitigation strategies. *IEEE Access*, 7:24703–24718, 2018.

[30] Marie-E. Godefroid, Ralf Plattfaut, and Björn Niehaves. IT Outside of the IT Department: Reviewing Lightweight IT in Times of Shadow IT and IT Consumerization. In Frederik Ahlemann, Reinhard Schütte, and Stefan Stieglitz, editors, *Innovation Through Information Systems*, pages 554–571, Cham, 2021. Springer International Publishing.

[31] Steffi Haag and Andreas Eckhardt. Shadow IT. *BUS INF SYST ENG*, 59(6):469–473, 2017.

[32] Julie M. Haney, Mary Theofanos, Yasemin Acar, and Sandra Spickard Prettyman. "we make it a big deal in the company": Security mindsets in organizations that develop cryptographic products. In William Enck and Adrienne Porter Felt, editors, *Proceedings of the 27th USENIX Security Symposium (USENIX Security)*. USENIX Association, 2018.

[33] Maximilian Häring, Eva Gerlitz, Matthew Smith, and Christian Tiefenau. Less about privacy: Revisiting a survey about the german covid-19 contact tracing app. In *Proceedings of the 2023 ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 2023.

[34] Jonas Hielscher and Simon Parkin. "What Keeps People Secure is That They Met The Security Team": Deconstructing Drivers And Goals of Organizational Security Awareness. In Patrick Gage Kelley Kelley and Apu Kapadia, editors, *Proceedings of the 20th Symposium on Usable Privacy and Security (SOUPS)*. USENIX Association, 2024.

[35] Ihab F Ilyas and Xu Chu. *Data cleaning*. Morgan & Claypool, 2019.

[36] Philip-Nicolas Johnson-Laird. *Mental models: Towards a cognitive science of language, inference, and consciousness.* Harvard University Press, 1983.

[37] Sebastian Käss, Marie Godefroid, Vincent Borghoff, Susanne Strahringer, Markus Westner, and Ralf Plattfaut. Towards a taxonomy of concepts describing it outside the it department. In *Proceedings of the 32nd Australasian Conference on Information Systems (ACIS)*, 2021.

[38] Michaela Kauer, Florian Kiesel, Felix Ueberschaer, Melanie Volkamer, and Ralph Bruder. The influence of trustworthiness of website layout on security perception of websites. In *Current issues in IT security*. Duncker & Humblot, 2012.

[39] Iacovos Kirlappos, Simon Parkin, and M Angela Sasse. Learning from "Shadow Security": Why understanding non-compliance provides the basis for effective Security. In *Proceedings of the 2014 Network and Distributed System Security (NDSS) Sympsoium*, 2014.

[40] Iacovos Kirlappos, Simon Parkin, and M Angela Sasse. "shadow security" as a tool for the learning organization. *ACM SIGCAS Computers and Society*, 45(1):29–37, 2015.

[41] Andreas Kopper and Markus Westner. Towards a taxonomy for shadow IT. In *Americas Conference on Information Systems*, 2016.

[42] Martin Kretzer and Alexander Maedche. Generativity of Business Intelligence Platforms: A Research Agenda Guided by Lessons from Shadow IT. In *Proc. of MKWI*, pages 207–229, 2014.

[43] Katharina Krombholz, Karoline Busse, Katharina Pfeffer, Matthew Smith, and Emanuel von Zezschwitz. "If HTTPS Were Secure, I Wouldn't Need 2FA" - End User and Administrator Mental Models of HTTPS. In *Proceedings of the 40th IEEE Symposium on Security & Privacy (S&P)*, pages 246–263. IEEE, 2019.

[44] Dong-Heon Austin Kwak, Xiao Ma, and Sumin Kim. When does social desirability become a problem? detection and reduction of social desirability bias in information systems research. *Information & Management*, 58(7):103500, 2021.

[45] Debin Liu, Farzaneh Asgharpour, and L Jean Camp. Risk communication in security using mental models. In Sven Dietrich and Rachna Dhamija, editors, *Proceedings of the 12th International Workshop on Usable Security (USEC)*, volume 4886 of *Lecture Notes in Computer Science*, pages 1–12. Springer, 2007.

[46] Alexandra Mai, Katharina Pfeffer, Matthias Gusenbauer, Edgar Weippl, and Katharina Krombholz. User Mental Models of Cryptocurrency Systems-A Grounded Theory Approach. In Heather Richter Lipford and Sonia Chiasson, editors, *Proceedings of the 16th Symposium on Usable Privacy and Security (SOUPS)*. USENIX Association, 2020.

[47] Gabriela Labres Mallmann, Aline de Vargas Pinto, and Antônio Carlos Gastaud Maçada. Shedding light on shadow it: Definition, related concepts, and consequences. In Paulo Silva Isabel Ramos, Rui Quaresma, editor, *Proceedings of the 18th Conference of the Portuguese Association for Information Systems*, pages 63–79. Springer, 2018.

[48] Gabriela Labres Mallmann, Aline de Vargas Pinto, and Antônio Carlos Gastaud Maçada. Shedding Light on Shadow IT: Definition, Related Concepts, and Consequences. In *Information Systems for Industry 4.0*, Lecturenotes in information systems and organisation, pages 63–79. Springer International Publishing, Cham, 2019.

[49] Heike Märki, Miriam Maas, Michaela Kauer-Franz, and Marius Oberle. Increasing software security by using mental models. In D Nicholson, editor, *Advances in Intelligent Systems and Computing*, pages 347–359. Springer, 2016.

[50] Frauke Mörike, Hannah L Spiehl, and Markus A Feufel. Workarounds in the shadow system: An ethnographic study of requirements for documentation and cooperation in a clinical advisory center. *Human factors*, 66(3):636–646, 2024.

[51] Maggie Oates, Yama Ahmadullah, Abigail Marsh, Chelse Swoopes, Shikun Zhang, Rebecca Balebako, and Lorrie Faith Cranor. Turtles, Locks, and Bathrooms: Understanding Mental Models of Privacy Through Illustration. In Rachel Greenstadt, Damon McCoy, and Carmela Troncoso, editors, *Proceedings of the 18th Privacy Enhancing Technologies Symposium (PETS)*, pages 5–32, Barcelona, Spain, 2018. De Gruyter Open.

[52] Selma Gomez Orr, Cyrus Jian Bonyadi, Enis Golaszewski, Alan T Sherman, Peter AH Peterson, Richard Forno, Sydney Johns, and Jimmy Rodriguez. Shadow it in higher education: Survey and case study for cybersecurity. *Cryptologia*, pages 1–65, 2022.

[53] Celeste Lyn Paul and Kirsten Whitley. A taxonomy of cyber awareness questions for the user-centered design of cyber situation awareness. In Louis Marinos and Ioannis Askoxylakis, editors, *Proceedings of the 1st International Conference on Human Aspects of Information Security, Privacy and Trust (HAS)*, pages 145–154. Springer, 1 edition, 2013.

[54] Robert W. Reeder, Adrienne Porter Felt, Sunny Consolvo, Nathan Malkin, Christopher Thompson, and Serge Egelman. An experience sampling study of user reactions to browser warnings in the field. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18. ACM, April 2018.

[55] Ita Ryan, Utz Roedig, and Klaas-Jan Stol. Understanding developer security archetypes. In *2021 IEEE/ACM 2nd International Workshop on Engineering and Cybersecurity of Critical Systems (EnCyCriS)*, pages 37–40. IEEE, 2021.

[56] Leonie Schaewitz, David Lakotta, M Angela Sasse, and Nikol Rummel. Peeking into the black box: Towards understanding user understanding of e2ee. In *Proceedings of the 2021 European Symposium on Usable Security*, pages 129–140, 2021.

[57] Koen Schoenmakers, Daniel Greene, Sarah Stutterheim, Herbert Lin, and Megan J Palmer. The security mindset: characteristics, development, and consequences. *Journal of Cybersecurity*, 9(1), 2023.

[58] Donald Sharpe. Chi-square test is statistically significant: Now what? *Practical Assessment, Research, and Evaluation*, 20(1):8, 2015.

[59] R John Simes. An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 1986.

[60] Eric Spero, Milica Stojmenovic, Zahra Hassanzadeh, Sonia Chiasson, and Robert Biddle. Mixed Pictures: Mental Models of Malware. In Ali Ghorbani, editor, *Proceedings of the 17th International Conference on Privacy, Security and Trust (PST)*, Fredericton, NB,Canada, 2019. IEEE.

[61] Nancy Staggers and Anthony F. Norcio. Mental models: concepts for human-computer interaction research. *International Journal of Man-machine studies*, 38(4):587–605, 1993.

[62] Mohammad Tahaei, Alisa Frik, and Kami Vaniea. Privacy champions in software teams: Understanding their motivations, strategies, and challenges. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.

[63] Jan Tolsdorf and Florian Dehling. In our employer we trust: mental models of office workers' privacy perceptions. In *Proceedings of the 24th Financial Cryptography and Data Security*, pages 122–136. Springer, 2020.

[64] Daniel W Turner. Qualitative interview design: A practical guide for novice investigators. *The Qualitative Report*, 15(3):754–760, 2010.

[65] Jan-Philip van Acken, Floris Jansen, Slinger Jansen, and Katsiaryna Labunets. Data underlying the research of case study of shadow it mindsets among corporate employees, 2024. Available at https://doi.org/10.24416/UU01-WEIBJU.

[66] Melanie Volkamer and Karen Renaud. Mental models– general introduction and review of their application to human-centred security. *Number Theory and Cryptography: Papers in Honor of Johannes Buchmann on the Occasion of His 60th Birthday*, pages 255–280, 2013.

[67] Rick Wash. Folk models of home computer security. In Lorrie Faith Cranor, editor, *Proceedings of the 6th Symposium on Usable Privacy and Security (SOUPS)*, pages 1–16. USENIX Association, 2010.

[68] Rick Wash and Emilee Rader. Influencing Mental Models of Security: A Research Agenda. In Sean Peisert, Richard L. Ford, Carrie Gates, and Cormac Herley, editors, *Proceedings of the 2011 New Security Paradigms Workshop (NSPW)*, pages 57–66. ACM, 2011.

[69] Rick Wash and Emilee Rader. Too Much Knowledge? Security Beliefs and Protective Behaviors Among United States Internet Users. In Lorrie Faith Cranor, Robert Biddle, and Sunny Consolvo, editors, *Proceedings of the 11th Symposium on Usable Privacy and Security (SOUPS)*, pages 309–325. USENIX Association, 2015.

[70] Martin S. White. Workarounds and shadow it - balancing innovation and risks. *Business Information Review*, 40(3):114–122, 2023.

[71] Wu, L Min, Robert C Miller, and Garfinkel. Do security toolbars actually prevent phishing attacks? In Robin Jeffries, editor, *Proceedings of the 2006 ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 601–610. ACM, 2006.

[72] Justin Wu and Daniel Zappala. When is a Tree Really a Truck? Exploring Mental Models of Encryption. In Sonia Chiasson and Rob Reeder, editors, *Proceedings of the 14th Symposium on Usable Privacy and Security (SOUPS)*, pages 395–409, Baltimore, MD, USA, 2018. USENIX Association.

[73] Stephan Zimmermann and Christopher Rentrop. Schatten-it. *HMD Praxis der Wirtschaftsinformatik*, 49(6):60–68, 2012.

Appendix A Codebook

Codebook levels — first part (Lvl 1 / Lvl 2 / Lvl 3, and Lvl 4). Columns: Client-facing (G), Support (G), Mngmnt (G), All groups (G, C).

Code (Lvl 1 / Lvl 2 / Lvl 3 / Lvl 4)	Client-facing G	Support G	Mngmnt G	All groups G	All groups C
Perspective					
Client laptop	15	0	0	15	7
Remote workspace	7	0	0	7	7
Client licenses	5	0	0	5	4
Shadow IT					
Definition	193	59	19	271	32
Unfamiliar	19	5	4	28	28
Familiar	13	5	4	22	22
Reasons for shadow IT	32	15	1	48	20
Need for functionality	7	3	0	10	10
Client requirement	8	0	0	8	4
Habit	2	3	0	8	8
Ease of use	4	2	0	6	5
Workaround	4	1	0	5	5
Insufficient standard	1	3	0	4	3
Time constraint	4	0	0	4	3
Financial feasability	0	2	0	2	2
Language barrier	1	0	0	1	1
Personal preference	1	1	1	3	3
Implications of shadow IT	53	17	10	80	32
Benefits	7	4	2	13	11
Efficiency	6	2	2	10	8
Cost efficiency	1	2	0	3	3
Risks	46	13	8	67	32
Data leak	16	4	3	23	23
Malware	11	2	1	14	14
Unauthorized access	8	2	0	10	10
Non-central governance	7	0	1	8	8
Reputation risk	2	2	1	5	5
Ransomware	1	1	1	3	3
Outdated software	1	1	0	2	2
Misinformation	0	1	1	2	2
Scenario	22	3	0	25	11
Approach IT team	6	2	0	8	8
Approach manager	5	0	0	5	5
Autonomous due dilligence	5	0	0	5	5
Approach client	3	0	0	3	3
Check internal store	1	1	0	2	2
Discuss with team	2	0	0	2	2
Types	68	19	4	91	28
Cloud services	26	8	1	35	17
Generative LLM	10	1	0	11	10
Online collaboration	6	2	0	8	5
Translate tools	3	4	0	7	7
Cloud storage	3	1	1	5	5
Browser extension	4	0	0	4	4
Self-installed applications	36	11	3	50	24
Streaming services	8	0	0	8	8

Codebook levels — second part (Lvl 1 / Lvl 2 / Lvl 3, and Lvl 4). Columns: Client-facing (G), Support (G), Mngmnt (G), All groups (G, C).

Code (Lvl 3 / Lvl 4)	Client-facing G	Support G	Mngmnt G	All groups G	All groups C
Code editor	5	2	0	7	7
Other browser	4	3	0	7	7
Network tool	3	1	0	4	4
Remote workspace	4	0	0	4	4
Mobile application	1	1	1	3	3
PDF reader	3	0	0	3	3
Conferencing tool	2	0	0	2	1
File reader	1	1	0	2	2
Automation software	0	1	1	2	2
Password manager	1	1	0	2	2
Version control app	1	1	0	2	2
Screencapture tool	1	0	1	2	1
Design tool	1	0	0	1	1
Virtual machine	1	0	0	1	1
Self-made solutions	4	0	0	4	2
Own software	2	0	0	2	1
External spreadsheet	1	0	0	1	1
System coupling	1	0	0	1	1
Private devices	2	0	0	2	2
Network device	1	0	0	1	1
Personal laptop	1	0	0	1	1
Policy & Awareness	63	15	9	87	32
Policy awareness	21	7	4	32	32
Unfamiliar	9	6	2	17	17
Familiar	12	1	2	15	15
Use of technology discussion	26	3	3	32	32
Formal	11	2	2	15	15
Informal	10	0	1	11	11
No discussion	5	1	0	6	6
Awareness perception	16	5	2	23	23
Reasonably well-informed	9	4	1	14	14
Well-informed	6	0	1	7	7
Not well-informed	1	1	0	2	2
Contradictions	11	7	2	20*	9
Mindsets					
Risk-averse	46	17	9	72	31
Consequence-avoidance orientations	25	6	4	35	23
Knowledge-based conservatism	9	6	2	17	17
Risk transfer mindset	7	0	1	8	8
Cautious seasoned judgment	6	0	0	6	6
Risk-taking	21	11	5	37	22
Common sense fallacy	8	1	2	11	11
Illusion of self sufficiency	3	3	0	6	6
Misguided sense of protection	3	2	1	6	6
Performance-driven rule bending	4	0	1	5	5
Longevity-based invincibility	2	2	1	5	5
Cost-driven compromise	1	3	0	4	4

C abbreviates coverage, the number of different experts mentioning the concept; G abbreviates grounding, the number of times a code occurred in total.. Note: * - the coverage for contradictions occurred at least twice per document; the category was used to code for the two statements that were perceived as being contradictory to each other.

Table 6: Interview Codebook with Coverage and Grounding

Appendix B   Informed Consent - Interviews

Informed consent

Information about the research

The interview you are asked to participate in is part of scientific research aiming to gain insights into the understanding and cybersecurity problems of shadow IT. Shadow IT is defined as "hardware, software, or services built, introduced, and/or used for the job without explicit approval or even knowledge of the organization" (Haag & Eckhardt, 2017).

How will the study be carried out?

The interview will take at maximum one hour, during which the researcher will ask questions in a semi-structured format. The interview will be recorded. After the recordings are transcribed, you will get the opportunity to remove any information from the text that should not be included in further analysis. Following the researchers' analysis of these transcripts, you will be asked to evaluate and add to a summary of the results that are based on the interviews. You will not be reimbursed for your participation in this study.

What will we do with your data?

During this interview, data about your experiences with shadow IT will be collected. Although the objectives and design of this study do not require specific personally identifiable information, the data collected should be considered as such. The interview will be recorded before it is transcribed. Interview recordings will be retained for up to six months until transcribed. The non-anonymised transcripts will only be processed by researchers who are collaborating in the study, or who are responsible for assessing its implementation. After analysis, the transcripts will be further anonymised as described in the next section. There are no specific increased privacy risks related to the nature of the collected personal data or the processing that the data will undergo. The data is stored and processed exclusively in the EU and all third party applications used have an appropriate data processing agreement with Utrecht University.

Processed data will be retained for at least 10 years for the purposes of research integrity. Before this archival, all personal information that can reasonably be traced back to you or your organization will have been removed or changed before the files are shared with other researchers or the results are made public. The researcher will keep a link that identifies you and your organization with the information, but this link will be kept secure and only available to the researcher. Any information that can identify you will remain confidential. The information in this study will only be used in ways that do not reveal who you are. You and your organization will not be named or identified in publications about this study or in documents shared with other researchers.

What are your rights?

Participation is voluntary. We are only allowed to collect your data for our study if you consent to this. If you decide not to participate, you do not have to take any further action. You do not need to sign anything. Nor are you required to explain why you do not want to participate. If you decide to participate, you can always change your mind and stop participating at any time, including during the study. You will even be able to withdraw your consent after you have participated. However, if you choose to do so, we will not be required to undo the processing of your data that has taken place up until that time. The research data we have obtained from you up until the time when you withdraw your consent will be erased.

Approval of this study

The Ethics and Privacy Quick Scan of the Utrecht University Research Institute of Information and Computing Sciences classified this research as low-risk and did not reveal any ethical problems for this research. If you have a complaint about the way this study is carried out, please send an email to the secretary of this Committee: etc-beta-geo@uu.nl. If you have any complaints or questions about the processing of personal data, please send an email to the Data Protection Officer of Utrecht University: privacy@uu.nl). The Data Protection Officer will also be able to assist you in exercising the rights you have under the GDPR. Please also be advised that you have the right to submit a complaint with the Dutch Data Protection Authority (https://www.autoriteitpersoonsgegevens.nl/en).

More information about this study?

In case you have additional questions, please contact Floris Jansen (researcher and data controller for the study) at f.j.jansen@students.uu.nl or Kate Labunets (project supervisor for the study) at k.labunets@uu.nl.

Haag, S., & Eckhardt, A. (2017). Shadow IT. Business & Information Systems Engineering, 59(6), 469–473.

I have read and understood the study information dated {date://CurrentDate/PT}, or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.

　　　Yes / No

I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.

　　　Yes / No

I understand that information I provide will be used for the report and publications in academic venues (like conferences or journals).

　　　Yes / No

I understand that personal information collected about me that can identify me, such as my name or email address, will not be shared beyond the study team.

　　　Yes / No

I additionally agree that my information can be quoted in research outputs

　　　Yes / No

I give additional permission for the pseudonymised interview transcript that I provide to be archived in UU's Yoda as open-access data so it can be used for future research and learning.

　　　Yes / No

Enter your name ......
Enter your email address......

Interview protocol

Pre recording Thank the interviewee for their willingness to participate, reiterate the research goals, and set expectations for the duration of the interview (around 30 mins) and the topics that will be covered.

1. Introduction
> Please state your rank, team, education and years of professional work experience
> What is the nature of your work? Do you do work in engagements?
>> If so, how many engagements have you done?
> What kind of work do you do?
> What kind of software do you need for your work tasks?
>> Have you ever needed special software for your clients?
>> some more indented text some more indented text

2. Understanding Shadow IT
> What is Shadow IT for you? (Could you please define what Shadow IT is?)
>> *If definition is known*: let the participant explain and introduce our definition
>> *If definition is unknown*: introduce our definition - "hardware, software, or services built, introduced, and/or used for the job without explicit approval or even knowledge of the organization"
> Introduce four types of shadow IT
>> Cloud services
>> Downloaded and install programs
>> Self-built solutions
>> Private devices

Occurrence of Shadow IT - Have you ever used?
> Cloud services - *for engagements? work tasks? personal use?*
> Downloaded and install programs - *for engagements? work tasks? personal use?*
> Self-built solutions - *for engagements? work tasks? personal use?*
> Private devices - *for engagements? work tasks? personal use?*

If a participant ever used a certain application -> Why those occurrences?
> Missing feature?
> Client request?
> Personal preference
> Time constraints?

4. Risks and implications of shadow IT?
> What do you think the risks are of the different types of shadow IT?
>> Risks for the user/participant?
>> Risks for your organization?
>> Risks for the client?
> What do you think are other implications of the different types of shadow IT?

5. Drawing exercise (only for client-specific software)
> Draw the process of the need to use client-specific applications.
>> So the client has asked you to work towards goal X, to do this you need an application that you do not have at the moment, how do you address this?

6. Policy and awareness
> Are you aware of the [organizational policy]?
>> If yes: could you quickly explain the policy?
>> If no: ask what their perception of the use of technology is within your organization.
>> Afterwards, explain the policy
> Have you discussed the use of technology amongst your team members?
> Do you feel you have been well informed about the use of technology?
>> (either through web learnings, your colleagues, training)
>> - do you think policy and awareness should do more?

7. Interview closing
> Would you like to add anything else?

Thank the interviewee for their time and explain further procedures of transcript review, member checking of codes, and sharing of results.

Of Mothers and Managers – The Effect of Videos Depicting Gender Stereotypes on Women and Men in the Security and Privacy Field

Nina Gerber
Technical University of Darmstadt
nina.gerber@tu-darmstadt.de

Alina Stöver
Technical University of Darmstadt
alina.stoever@tu-darmstadt.de

Peter Mayer
University of Southern Denmark
Karlsruhe Institute of Technology
mayer@imada.sdu.dk

Abstract

Gender imbalances are prevalent in computer science and the security and privacy (S&P) field in particular, giving rise to gender stereotypes. The existence of such stereotypes might elicit the *stereotype threat effect* well-known from research in math settings: mere exposure to stereotypes can decrease the performance in and attitude towards specific fields. In this work, we investigate whether the stereotype threat effect influences women and men in the S&P field. We conducted an online experiment with multiple groups to explore whether videos that depict and counteract gender stereotypes influence S&P attitudes and intentions (RQ1), and (self-assessed) S&P knowledge (RQ2). We find overall little evidence for the stereotype threat effect, but our results show that women in the condition actively counteracting gender stereotypes report a higher interest in preventing hacker access to their devices than women in the stereotype conditions. In addition, we find that men score higher than women in a variety of self-report measures, except for security and privacy concerns. These results indicate that stereotypes might need to be addressed early on to prevent stereotypes from becoming social norms and a self-fulfilling prophecy of gender imbalance in the S&P field.

1 Introduction

Computer science in general and the security and privacy field in particular are among the fields where gender imbalances are the most pronounced [5, 12, 52]. In fact, skills required for computer science are often perceived as incompatible with female gender roles [9]. Luckily, a variety of successful programs are trying to counter that imbalance [11, 19, 57]. Yet, research has shown that the security and privacy field is riddled with negative stereotypes [70].

These stereotypes might elicit in women trying to enter the security and privacy field what is known as stereotype threat. This effect has been well-documented in the field of mathematics [61]: when individuals are exposed to depictions or descriptions of stereotypes that target them, it can affect the objective performance and interest in the respective domain of these individuals. For instance, in [16] exposure to gender stereotypes portraying commercials decreased women's performance in a math test (despite the stereotypes not being math performance-related), while women who saw counter-stereotypic commercials performed as well as men did in the same test. The stereotype threat effect has been shown to affect individuals targeted by a wide variety of stereotypes, such as ethnicity (e.g., [3]) or gender (e.g., [47]).

In this work, we investigate whether stereotypes portrayed in commercials videos can elicit the stereotype threat effect and affect security and privacy (S&P) attitudes, and (self-assessed) S&P knowledge in the same manner as they can in the mathematics context. To that end, we conducted a 4x2-between-subject online randomized controlled trial experiment with $N = 959$ participants. We tested a variety of security and privacy aspects – including security attitude, security behavior intention, technological affinity, and privacy concerns – across four experimental conditions (stereotype women, stereotype men, non-stereotype, control) and across men and women.

Specifically, we investigated the following two research questions:

RQ1: *Do videos that depict gender stereotypes influence S&P attitudes and intentions?*
Women in the non-stereotype condition reported more interest in preventing hackers from getting access to their devices. Men overall scored higher on the measured scales, except for concerns where women scored higher.

RQ2: *Do videos that depict gender stereotypes influence (self-assessed) S&P knowledge?*
Women in the group with videos depicting women in stereotypical settings reported higher levels of computer security knowledge than women in the other groups. Men performed better in terms of the S&P knowledge metrics than women.

Our paper makes the following contributions:

- We show that the stereotype threat effect does not seem to affect the S&P domain in the same way as in [16], highlighting the need to identify alternative factors influencing gender imbalances in S&P.
- We validate and extend prior work by showing that there exist differences between men and women regarding their S&P attitudes, intentions, (self-assessed) knowledge, and interest.
- We contextualize our findings in the related work and discuss implications for priming studies.

2   Related Work

2.1   Stereotype Threat

Stereotype threat is a psychological phenomenon which inhibits the performance of individuals in real-life situations when they are confronted with descriptions or depictions of negative stereotypes pertaining to the performance of specific groups of people they are part of [54]. Among the first cases where this effect was observed was the performance of women in math tests [61]. Their performance could be manipulated by either telling them that math performance was affected by gender, in which case they performed worse than similarly skilled men, or that it was not affected by gender, in which case they performed similarly to the men [61]. The same difference in performance could be observed when women were just told their performance would be checked in a math test versus them being made aware of the stereotype threat effect [37]. In essence it induces anxiety that impairs short-term academic performance [54]. It has since been shown to impact a wide variety of individuals from various backgrounds that are affected by negative performance stereotypes (e.g, [3, 47, 50, 60]). The effect has been shown to appear as early as elementary and middle school [25, 54].

Effective countermeasures to the stereotype threat effect include explicit communication contradicting the stereotype [48], describing the skill required for a particular task as malleable [3, 25], re-framing it as a challenge [2], or re-framing it as consequence of a specific situation that was possible to overcome (such as difficulties in math after changing from elementary to middle school) [25]. Offering support to affected individuals has also been shown to be an effective counter-measure [11, 19, 57].

The stereotype threat effect and whether it is possible to observe this effect in the security and privacy domain is the main subject of this work.

2.2   Gender Stereotypes in Advertising

Advertisements are strongly biased in terms of gender representation [49]. Men are far more present with more screen time and more voiceovers than women. Women are also still depicted in traditional gender roles, e.g., as housewives, as opposed to men who are depicted as independent or with physical activity. Furthermore, the sexualization of both genders is increasing with a steeper increase in the depictions of women [49], despite the fact that a recent meta-review [36] found sexualized depictions in advertisements has no effect on purchase intention and there is even a small negative effect on brand attitude. In fact, it was found that these portrayals are perceived as not representing contemporary society [28].

Critically, the perpetuation of gender stereotypes can happen early [38] in a person's life and in advertisements indeed increasingly affects children [42, 53, 55]. Stereotypical colors are used to indicate whether a product is meant for girls or boys and toys aimed at girls put a focus on appearance, nurturing, and cooperation while toys aimed at boys put a focus on competition, independence, and physical activity [4, 14]. Due to the pervasive nature of advertisements in our society, these stereotypical depictions have the potential of influencing a wide audience. They have been found to elicit stereotype threats relating to math problems among women [16] and they might shape children's understanding of gender [4, 7] and in turn their interests and behavior [51].

Therefore, advertisement videos represented the ideal choice for our study. If stereotypical depictions in advertisements can influence adults' and children's interest in security and privacy topics this could have detrimental effects on the respective protections people employ.

2.3   Gender Stereotypes in the STEM Field

Research has indicated that some differences between men and women exist in the STEM field in general and in the security and privacy domain in particular. When considering the wider STEM field, it has been found that women are more likely to experience a lack of support [11] and systematic support structures have a positive impact on women staying in their chosen discipline [57]. Whether women choose a major in the STEM field is also influenced by stereotypes. Particularly, nerd-genius stereotypes have been shown to negatively impact women's STEM identity [62]. More specifically for computer science in the STEM field, it was found that traits needed for computer science are perceived as incompatible with female gender roles [9]. When these traits were described as outdated stereotypes, women expressed more interest in computer science.

When looking at gender imbalances in the domain of security and privacy, women seem to have higher confidentiality and integrity concerns than men [41] and while women also feel more negatively about tracking, they are less likely to employ protective actions [13]. Interestingly, there also seems to be some evidence indicating that women might be at higher risk from cyber threats. For instance, they seem to be less aware of data breaches in which their data was involved [46] and more susceptible to phishing [59]. This issue is conflated by a wide array of negative stereotypes towards women in the security and privacy domain [70].

Considering this evidence on gender stereotypes, the reason for performance differences is likely to be connected to stereotype threat as has been well documented in other domains. Therefore, we chose to investigate the stereotype threat effect in the domain of security and privacy.

3 Methodology

We conducted an online experiment following a between-subject design to explore whether videos that depict and counter gender stereotypes influence security and privacy (S&P) attitudes and intentions (RQ1), and (self-assessed) S&P knowledge (RQ2).

3.1 Selection of Videos and Study Conditions

We used videos in our study to elicit gender stereotypes in our participants which we identified in a multi-stepped procedure. As already outlined in section 2, we decided to base our investigation on advertisement videos since it has already been shown that they can elicit the stereotype threat effect [16]. Specifically, we used commercials as treatment (following similar work by Davies et al. [16]), as opposed to text instructions (used by e.g., Johns et al. [37]) since we sought to understand whether real-world commercials have the potential to negatively impact women in the STEM field.

Step 1: Initial Search. To identify suitable videos, we performed a search on the YouTube video streaming platform with search terms informed by the related work: one of either "ad", "advertisement" or "commercial" combined with one of "baby formula", "detergent", "stroller", "car", "e-car", "tech", "insurance", "bank", "smart home", "stem", "science", "space", "engineering", and "cosmetics".

From the videos we found, we decided to choose videos fitting four study conditions. Firstly we chose videos that depicted women in stereotypical situations (*Stereotype women* condition), e.g., as mothers or spouses, akin to the work in [16]. Secondly, we searched for videos that depicted men (but not women) as stereotypical representatives of the engineering and science domains (*Stereotype men* condition). The message to women in this condition might be that men

rather than women are typically working there, playing into perceived social norms [9]. Thirdly, we chose videos that used non-stereotypical representations of women (*Non-stereotype* condition). Specifically, we chose videos promoting campaigns for women in STEM and computer science. Since these videos are explicitly created with countering stereotypes in mind, we felt it was the strongest opposite of the two stereotype conditions. Last but not least, we chose neutral videos, as the baseline for our comparison (*Control* condition). Specifically, we chose non-anthropomorphized depictions of animals (e.g., horses running across mountain landscapes)[1]. We selected several videos for each of these conditions. The most suitable two for each condition were selected in a pre-study as outlined below.

Step 2: Pre-Study to Select Most Suitable Videos. To identify the most suitable video for each condition (stereotype, non-stereotype, control), we conducted an online survey as pre-study. The survey had the participants watch several of the candidate videos in a randomized order and for each video rate to what degree the shown video includes several stereotypes. The full questions can be found as an online appendix on GitHub[2]. We recruited $n = 92$ participants off the Prolific platform. They were compensated with $3.76. The mean duration of the pre-study was 17:15 minutes.

Based on our results we identified two videos for each condition that would be shown in a random order in the main study. Specifically, these videos were:

- [*Stereotype women* condition] For the first stereotype condition, we chose two videos that depicted women (but not men) in homemaker settings. The first video depicts women as taking care of babies. The second video depicts a woman in a family setting, preparing food for the family and doing the laundry. These two videos were rated highly by our pre-study participants in terms of stereotypical depictions of women and portrayals of women primarily as parent/spouse, as opposed to neutral/low ratings for the other stereotypes which together with their content made them an ideal choice: *McDonald's - Stroller moments*[3] and *Tide Laundry Detergent - Muffins*[4].

- [*Stereotype men* condition] For the second of our stereotype conditions, we chose videos that depicted men (but not women) in engineering and science settings. The first video depicted men testing a car as engineers. The second one shows many different individuals in a variety of situations, where men are frequently depicted

---

[1]Note: While not apparent from the search terms, we found enough of these videos, e.g. bank ads made a great source for these.

[2]https://github.com/petermayer/snp-gender-stereotype-threat-priming-study

[3]https://www.youtube.com/watch?v=vkQ2dkqDFd0

[4]https://www.youtube.com/watch?v=l0cAK9ouRXU

Figure 1: Study procedure.

as scientists or engineers and women in family settings. These two videos were rated highly by our participants in terms of men being portrayed as engineers/scientists, as opposed to low/neutral ratings for the other stereotypes: *Fiat - Fiat 500 S*[5] and *DBS - Live more*[6].

- **[*Non-stereotype* condition]** As non-stereotype condition, we chose videos that depicted women in engineering and computer science settings. The following two campaign ads with *STEM* themes were rated highly by our participants in terms of women being portrayed as engineers/scientists, opposed to low/neutral ratings for the other stereotypes and due to their nature as campaign ads were specifically non-stereotype: *Kode With Klossy x #SheCanSTEM*[7] and *Dare to STEM*[8].

- **[*Control* condition]** For the control condition, we selected the following two videos due to their overall low/neutral ratings for all stereotypes and their content based on non-anthropomorphized depictions of animals: *Lloyds Bank – Epic Journey*[9] and *Mercedes-Benz - Chicken*[10].

We included questions with the same ratings as used in the pre-study questionnaire as manipulation checks in the main study, which confirmed the ratings from the pre-study.

3.2 Study Procedure

The participants were randomly assigned to one of the four video priming conditions. After consenting to the study, they were shown two advertisement videos, which were selected based on their video priming group assignment and directly embedded in the survey.

To investigate RQ1 (S&P Attitudes and Intentions), they were then asked to answer the SA-13 questionnaire [20] to capture their *security attitude*, the SeBIS scale [18] to capture their *security behavioral intention*, the ATI scale [21] to measure their *technological affinity*, the IUIPC-8 questionnaire [27, 44] to capture *privacy concerns*, and 12 statements

on a 5-point Likert-like scale taken from Story et al. [63] to assess their *interest in preventing various S&P risk scenarios*, such as hackers gaining device access.

In addition, the participants were asked to complete the following scales and items to investigate RQ2 ((self-assessed) S&P Knowledge): the Technical Knowledge of Privacy Tools Scale [39] consisting of six true/false/I'm not sure items to measure their *technical knowledge of privacy tools*, the OPLIS Technical scale [67], including five multiple choice questions to capture their *technical privacy literacy*, the Internet Know-How Self Report Scale [39] to measure *familiarity with internet tools and concepts*, five items proposed by Sawaya et al. [58] to assess *self-confidence in security knowledge*, three items taken from Bermejo Fernandez et al. [6] to measure *general technical knowledge*, *computer security knowledge*, and *privacy knowledge*, and a self-constructed multiple choice question asking for *S&P skills*.

The participants were then asked to complete the Social Identities and Attitudes Scale (SIAS) [54] to measure their *identification with their respective gender* and the Ambivalent Sexism Inventory (ASI) [23] that captures *sexism*, followed by demographic questions and the option to make a comment to the study. Finally, we asked them to rate the videos they had seen with regards to reflecting general gender stereotypes, and specific gender stereotypes related to the video priming groups, i.e., displaying women and men as engineers/scientists, parent/spouse, and proficient in IT, as a *manipulation check*. After that, we *debriefed* them about the study purpose and explained that the videos they had seen might have contained inappropriate stereotypes, thanked them again for their participation, and redirected them to Prolific. Two attention check questions were included in the study. The final survey can be found as an online appendix on GitHub[11].

On average, it took 16:07 minutes (SD=5:17, Med=17:10) to finish the study. The study was pilot tested with 12 participants recruited via Prolific, who voiced no concerns or needs for adjustments.

3.3 Data Analysis

We conducted a set of one-way ANOVAs to compare S&P attitudes and intentions between the four video priming groups

---

[5] https://www.youtube.com/watch?v=3YBhftZSlbM
[6] https://www.youtube.com/watch?v=BJurmEJ6dNk
[7] https://www.youtube.com/watch?v=WE1r0vY95fU
[8] https://www.youtube.com/watch?v=0o9DeumoTkw
[9] https://www.youtube.com/watch?v=Rkz6X5VrRBU
[10] https://www.youtube.com/watch?v=nLwML2PagbY

[11] https://github.com/petermayer/snp-gender-stereotype-threat-priming-study

Table 1: Study participants' demographics.

	Women	Men
Age		
18-25	10.8%	11.8%
26-35	25.8%	35.9%
36-45	20.4%	24.8%
46-55	18.7%	15.4%
56-65	15.7%	9.4%
66-75	7.7%	2.6%
>76	0.9%	0.2%
Education		
High School Diploma	34.2%	32.9%
Bachelor's Degree	40.2%	41.7%
Master's Degree	12.9%	17.1%
Ph.D. or higher	1.9%	3.0%
Other	10.3%	5.3%
Prefer not to say	0.4%	/
Occupation		
Employed	55.1%	70.3%
Self-employed	13.8%	10.3%
Unemployed	5.8%	8.3%
Student	4.1%	5.1%
Retired	9.5%	3.8%
Homemaker	8.6%	1.1%
Other	2.8%	1.1%
Prefer not to say	0.4%	/
IT Experience		
Yes	19.1%	43.2%
No	80.9%	55.3%
Prefer not to say	0.2%	1.5%
	M (SD)	*M (SD)*
Hostile Sexism	2.74 (0.77)	3.07 (0.87)
Benevolent Sexism	3.12 (0.81)	3.25 (0.80)
Gender Identification	4.85 (1.32)	4.42 (1.41)

(RQ1) and unpaired t-tests to compare S&P attitudes and intentions between women and men, since these were measured with validated scales and met all assumptions for parametric testing. In case that homogeneity of variances was not given, we used Welch's ANOVA and Welch's t-test instead. Yet, interest in preventing S&P risk scenarios was only captured with single items and thus analyzed with the non-parametric Kruskal-Wallis tests to analyze the effects of the video priming and Wilcoxon rank-sum tests to analyze gender differences, following recommendations for Likert scales and single items in Likert response format [8].

Further, we conducted Kruskal-Wallis tests to analyze the effects of the video priming and Wilcoxon rank-sum tests to analyze gender differences in terms of S&P (self-assessed) knowledge (RQ2), since knowledge test performance is as-

sumed to be ordinal rather than metric, and self-assessed knowledge was measured with single items.

We decided to analyze the video priming effects for women and men separately, as the video priming displaying gender stereotypes can affect both groups differently [16]. For all post-hoc tests, we used Bonferroni-Holm-corrected alpha-levels. Since we have four video priming conditions, the Bonferroni-Holm-corrected alpha-levels are .05, .025, .0167, and .0125 respectively.

We performed an a priori power analysis to calculate the number of participants needed to detect a medium effect ($f = 0.25$; $d = 0.5$) with two-tailed testing ($\beta = 0.95$ and $\alpha = .05$). The analysis indicated a required total sample size of 840 participants for analyzing the parametric data, and a required total sample size of 880 for analyzing the non-parametric data, each including the potential post-hoc tests.

3.4 Recruitment and Participants

We used Prolific to recruit a sample of participants from the U.S., which was balanced regarding sex. Still, we made sure to include participants from all genders using the prescreen function in Prolific. Participants received an hourly wage of $14.38 for their participation. A total of 979 participants completed the questionnaire, of whom 20 were excluded due to failing at least one attention check. Of the remaining 959 participants, 465 identified as women, 468 as men, 14 as non-binary, and one each as trans man, trans women, trans masculine, demigirl, and "born with vagina". We focused our analysis on the participants identifying as either women or men, as we were interested in gender-specific effects and had only sufficient sample sizes for those two gender groups. Our final sample thus included 933 participants, which still well exceeds the required sample size of 880. For the participants' demographics, the reader is referred to Table 1, and to Table 11 in the appendix for a detailed breakdown of the demographics.

3.5 Ethics

The study received IRB approval. All participants provided consent for their participation and for their data being used prior to the study. They were told that they would see advertisement videos embedded in the survey via YouTube and that they therefore also had to consent to YouTube's terms and conditions by taking part in the study. Further, they were informed that they could quit the study at any time, in which case all data collected so far would be deleted. For this, participants could simply close the survey or click on a button labeled "Leave and delete my data". In addition, participants who wished to withdraw from the study after completion could contact us via email or the Prolific platform. At the end of the survey, we included a debriefing text to inform the participants about the research questions, highlighted that the commercials they had seen might have contained stereotypical gender representa-

tions, and pointed out that these stereotypical representations do not necessarily correspond to the truth.

3.6 Limitations

Like most experimental studies, our study is subject to several limitations.

First, the video selection relied on the search function on YouTube, which is highly personalized to users through intransparent algorithms [24, 66]. While we tried to minimize the influence of this personalization by searching in fresh browser sessions in private/incognito mode in different browsers, there is no way for us to guarantee that searches performed with, e.g., other browsers and OSes, would not have yielded additional search results. However, we found enough suitable videos that matched our selection criteria (content and stereotype ratings) in our pre-study, which makes us believe that additional search results would not have influenced our findings substantially beyond a negligible extent.

We only included a selection of stereotypes related to traditional roles of women and the traditional dominance of men in the technical field. A broader focus might have yielded further results for other stereotypes. Also, some of the videos were aired several years ago and would perhaps no longer be broadcast in this form today, and the STEM campaign videos are targeted at young women, while our sample included women and men of all ages. Still, the pre-study and the manipulation check confirmed that the videos successfully transferred the intended stereotypes and counter-stereotypes as needed to explore our research questions.

Second, we used Prolific for recruitment, which has been found generally representative for the U.S. population with regards to security and privacy experiences, perceptions, and beliefs, but not knowledge and self-reported behavior, particularly in terms of on social media use [1, 65]. As a result, our sample might perform better in the privacy knowledge tests and report security-related or privacy-related actions that may not reflect those of the general U.S. population. In addition, we only considered participants residing in the U.S. to avoid cultural differences in the groups as unintended additional influence besides the video priming. Hence, further research is needed to explore how gender stereotypes affect women and men with varying cultural backgrounds. We further focused our analysis on participants identifying as women or men, as these were the only gender groups with sufficiently large sample sizes for statistical analysis. Still, we acknowledge that there are multiple other gender groups such as non-binary, and highlight the importance of considering participants from those groups in future research, especially with regards to gender stereotypes. Finding ways to recruit participants with other gender roles in sufficient sample sizes and incorporate them in the analyses is an important line of future work.

Third, although we checked how strong our participants identified with their gender, we did not ask about their identifi-

cation with the depicted gender stereotypes. Yet, participants who identify strongly with the stereotypes presented, may react more strongly to them than participants who identify less with those stereotypes.

Fourth, the videos depicting and counteracting gender stereotypes might have affected women's and men's responses differently, exaggerating or understating existing gender differences. Further, the men and women in our sample reported considerably different levels of IT experience. While these differences might reflect actual gender imbalances in this field, it is also possible that the men in our sample were more and the women less tech-savvy compared to the general U.S. population.

4 Results

4.1 RQ1: S&P Attitudes and Intentions

Figure 2 shows women's and men's security and privacy (S&P) attitudes and intentions across the video priming groups (RQ1). For the detailed test results, the reader is referred to the appendix.

Security Attitude. Across the four video priming groups, women and men both reported on average moderate levels of security attitude (measured with the SA-13 questionnaire [20]). We did not find significant differences between the four video priming groups for women or men.

Security Behavior Intention. On average, women and men in all four video priming groups reported rather high levels of security behavior intention in terms of device securement, password generation, proactive awareness, and updating (measured with the SeBIS scale [18]). The analysis results did not indicate significant differences between the four video priming groups for women and men.

Technological Affinity. On average, women across the four video priming groups reported low to medium levels of technological affinity, whereas men in all four video priming groups reported medium to high levels of technological affinity (measured with the ATI scale [21]). Two one-way ANOVAs did not indicate significant differences between the four video priming groups for women or men.

Privacy Concerns. Both women and men in all four video priming groups reported high levels of privacy concerns (measured with the IUIPC-8 questionnaire [27, 44]). A set of one-way ANOVAs did not indicate significant differences between the four video priming groups for women or men.
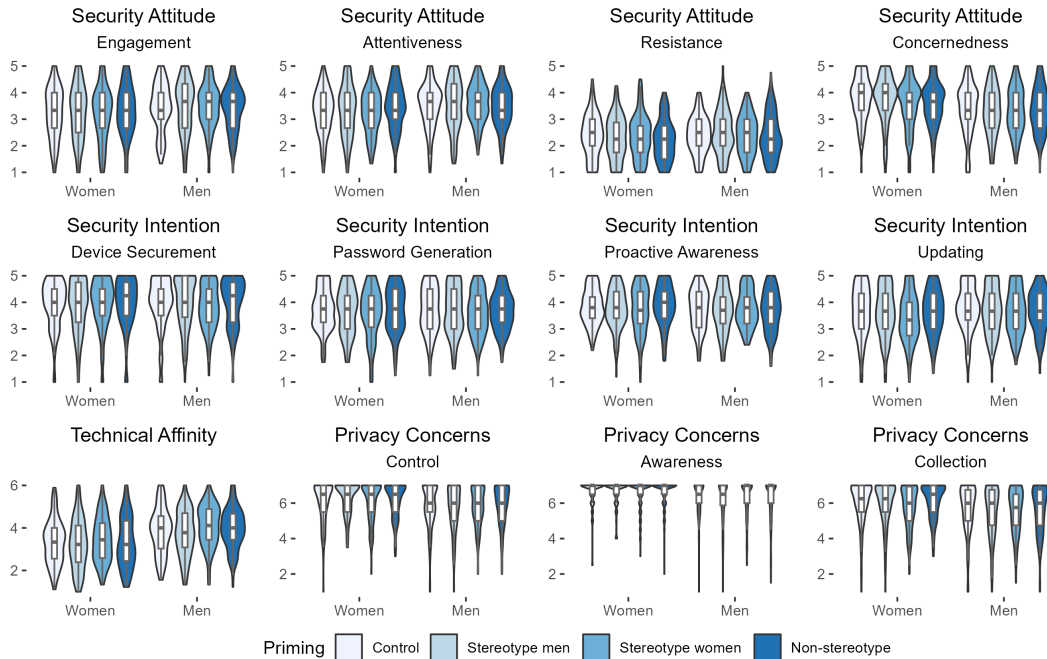
Figure 2: Violin and box plots showing the results for security attitude [20] (1=strongly disagree, 5=strongly agree), security behavioral intention [18] (1=never, 5=always), technological affinity [21] (1=strongly disagree, 6=strongly agree), and privacy concerns [27, 44] (1=strongly disagree, 7=strongly agree). The width of the curves represents the frequency of data points in each region, i.e., the wider the curve gets at a certain value, the more participants have indicated this value. The central line in the box plots marks the median, whereas the boxes indicate the central 50% of the data.
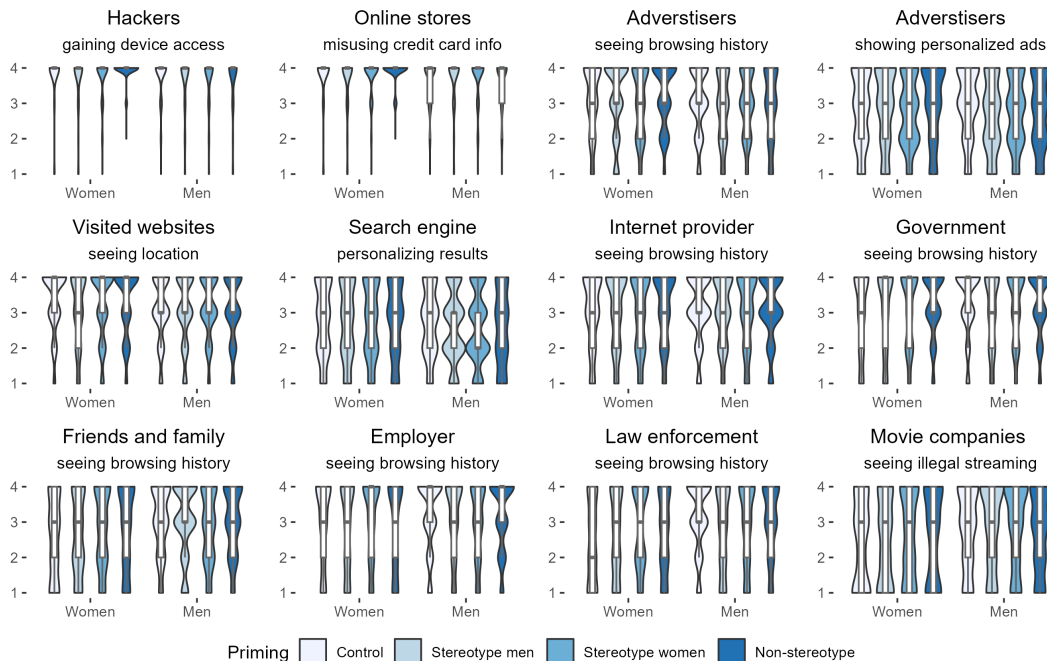
Figure 3: Results for interest in preventing various S&P risk scenarios using 12 statements taken from Story et al. [63] (1=not at all interested, 4=very interested).

Interest in Preventing S&P Risk Scenarios. Figure 3 shows women's and men's interest in preventing various scenarios describing security incidents or privacy infringements (measured with 12 Likert-like items taken from Story et al. [63]). Due to the ordinal scale level of the data, we calculated non-parametric tests (the detailed test results can be found in the appendix).

On average, women and men in all four video priming groups reported a great interest in preventing hacker access to their device, and misuse of their credit card information by online stores. Further, they reported a moderate to great interest in preventing advertisers, law enforcement, the government, their Internet provider, friends and family, and employer seeing their browsing history, advertisers showing personalized ads, websites they visit seeing their physical location, search engines showing personalized results, and movie companies seeing illegal movie streaming.

A Kruskal-Wallis test revealed significant differences in women's interest in preventing hackers from gaining access to their device between the four video priming groups ($\chi^2(3)$=11.696, p=.009, η^2=.019). Pairwise comparisons using Wilcoxon rank-sum tests with Bonferroni-Holm corrections of the alpha-level showed that women in the *Non-stereotype* condition reported significantly higher levels of interest to prevent hacker access to their device than women in the *Stereotype women* condition (Z=-3.008, p=.003, r=.197). Likewise, women in the *Non-stereotype* condition reported significantly higher levels of interest in preventing hacker access to their device than women in the *Stereotype men* condition (Z=-2.988, p=.003, r=.197), both indicating a small effect [10]. We could not replicate these effects for men. A set of further Kruskal-Wallis tests did not show significant differences for the other S&P risk scenarios between the four video priming groups for both women and men.

Summary. Women in the *Non-stereotype* condition reported more interest to prevent hacker access to their devices than women in the two video priming groups displaying gender stereotypes.

4.2 RQ2: Knowledge Test Performance and Self-Assessed Knowledge

Figure 4 shows women's and men's privacy knowledge test performance as well as self-assessed S&P knowledge and skills across the video priming groups (RQ2). The detailed test results can be found in the appendix.

Technical Knowledge of Privacy Tools. On average, women in all four video groups had rather little technical knowledge of privacy tools (measured with the Technical Knowledge of Privacy Tools Scale [39]), while men in all four groups had moderate knowledge. The analysis results did not indicate significant differences between the four video priming groups.

Technical Privacy Literacy. Both women and men in all four video groups scored rather high in terms of technical online privacy literacy (measured with the OPLIS Technical scale [67]). The analysis results did not indicate significant differences between the four video priming groups.

Familiarity with Internet Tools and Concepts. Both women and men in all four video priming groups reported medium levels of familiarity with Internet tools and concepts (measured with the Internet Know-How Self Report Scale [39]). We did not find significant differences between the four video priming groups.

Self-Confidence in Security Knowledge. Both women and men in all four video priming groups reported a medium level of confidence in their security knowledge (measured with 5 items proposed by Sawaya et al. [58]). The analysis results did not indicate significant differences between the four video priming groups.

Self-Assessed Technical Knowledge. Both women and men in all four video priming groups reported medium levels of general technical knowledge, computer security knowledge, and privacy knowledge (measured each with a Likert-item taken from Bermejo Fernandez et al. [6]). The analysis results revealed significant differences in self-assessed computer security knowledge between the four video priming groups for women ($\chi^2(3)$=8.570, p=.036, η^2=.012). Pairwise comparisons using Wilcoxon rank-sum tests with Bonferroni-Holm corrections of the alpha-level showed that participants in the *Stereotype women* condition reported higher levels of computer security knowledge than participants in the *Stereotype men* condition (Z=-2.693, p=.007, r=.176), indicating a small effect [10]. These results could not be replicated for men.

Self-Assessed S&P Skills. Roughly the same number of women described their S&P skills (measured with a self-constructed multiple choice question) as novice or competent and only a very small proportion as expert in all four video priming groups. In all four video groups, most men described their skills as competent, followed by novice and expert. Using Kruskal-Wallis tests, we did not find significant differences between the four video priming groups.

Summary. Women in the *Stereotype women* condition reported higher levels of computer security knowledge than women in the *Stereotype men* condition.
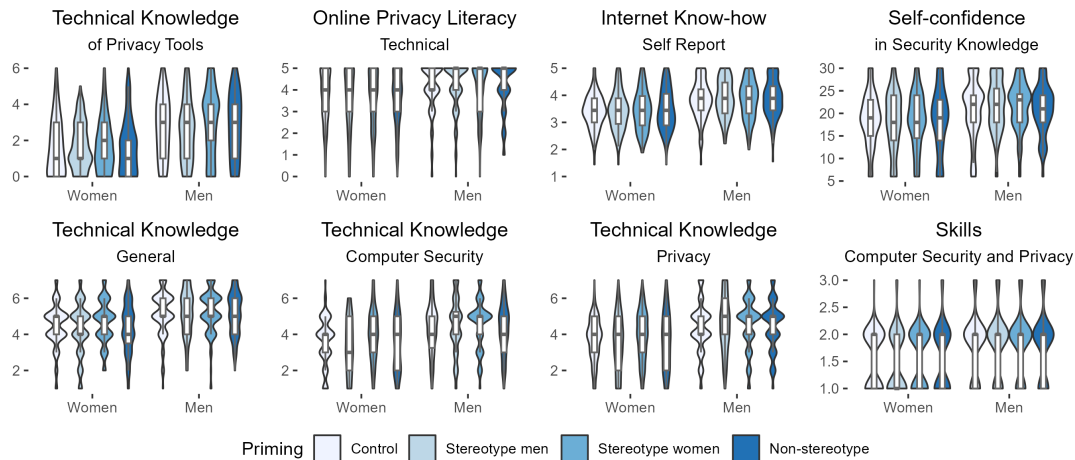
Figure 4: Results for technical knowledge of privacy tools [39] (1=low, 6=high), technical privacy literacy [67] (1=low, 5=high), familiarity with internet tools and concepts [39] (1=I've never heard of this, 5=I know very well how this works), self-confidence in security knowledge [58] (sum of 6 items from 1=strongly disagree to 5=strongly agree), general technical knowledge, computer security knowledge, and privacy knowledge [6] (1=low, 7=high), and S&P skills (1=novice, 2=competent, 3=expert).

4.3 Gender Effects

We further calculated unpaired t-tests (and Welch's t-tests, respectively, in case that homogeneity of variance was not given) and Wilcoxon rank-sum tests to analyze gender differences. For the detailed test results, the reader is referred to the appendix.

S&P Attitudes and Intentions. The results of the unpaired t-tests and Welch's t-tests indicated small significant differences between women and men for all four scales of the SA-13 questionnaire [20] measuring security attitude, with men reporting higher levels of security engagement (t(931)=-3.256, p=.001, d=-0.213), attentiveness (t(931)=-3.289, p=.001, d=-0.215), and resistance (t(931)=-1.969, p=.049, d=-0.129) than women, while women reported higher levels of security concernedness than men (t(931)=5.795, p<.001, d=0.379). Our results did not indicate significant differences between women's and men's security behavior intention as measured with the SeBIS scale [18]. Using Welch's t-tests, we further found significantly higher levels of self-reported technological affinity for men than for women with a medium effect size (t(923.846)=-8.211, p<.001, d=-0.538), and small significant gender differences for all three scales of the IUIPC-8 questionnaire [27, 44] measuring privacy concerns, with women indicating higher levels of control (t(917.367)=4.662, p<.001, d=0.305), awareness (t(863.300)=4.784, p<.001, d=0.313), and collection concerns (t(914.466)=5.009, p<.001, d=0.328) than men.

With regards to the different S&P risk scenarios, a set of Wilcoxon rank-sum tests showed that women reported significantly higher levels of interest than men to prevent hacker access to their device (Z=-3.046, p=002, r=0.100), misuse of credit card information by online stores (Z=-2.992,

p=.003, r=0.098), advertisers seeing their browsing history (Z=-2.730, p=.006, r=0.089), visited websites from seeing their location (Z=-2.930, p=.003, r=0.096), and receiving personalized search results based on their browsing history (Z=-2.464, p=.014, r=0.081).

Men, on the other hand, reported significantly higher levels of interest than women in preventing friends and family from seeing their browsing history (Z=-3.231, p=.001, r=0.106), along with law enforcement seeing this browsing history (Z=-3.706, p<.001, r=0.121), and movie companies seeing illegal movie streaming (Z=-3.624, p<.001, r=0.119), with small effect sizes for all gender differences.

Knowledge Test Performance and Self-Assessed Knowledge. Using Wilcoxon rank-sum tests, we found higher levels of technical knowledge of privacy tools (measured with the Technical Knowledge of Privacy Tools Scale [39]) for men than for women with a small to medium effect size (Z=-9.471, p<.001, r=0.310). In addition, we found significantly higher levels of technical online privacy literacy (measured with the OPLIS Technical scale [67]) for men than for women with a small effect size (Z=-6.003, p<.001, r=0.197).

Men also reported significantly higher levels of familiarity with Internet tools and concepts (measured with the Internet Know-How Self Report Scale [39]) than women with a small effect size (Z=-9.039, p<.001, r=0.296). Likewise, men reported slightly higher levels of confidence in their security knowledge (measured with 5 items proposed by Sawaya et al. [58]) than women. This difference was statistically significant with a small effect size (Z=-6.741, p<.001, r=0.221). We further found significantly higher values of self-assessed general technical knowledge for men than for women (Z=-9.009, p<.001, r=0.295), along with higher values of computer se-

curity knowledge (Z=-8.141, p<.001, r=0.267), and privacy knowledge (Z=-8.183, p<.001, r=0.268; measured each with a Likert-item taken from Bermejo Fernandez et al. [6]), all with a small effect size. In addition, the analysis results indicated significantly higher levels of self-assessed S&P skills for men than for women with a small effect size (Z=-6.525, p<.001, r=0.214). As IT experience might influence self-assessed S&P skills, we further performed an ordinal regression analysis, which confirmed an effect of IT experience on self-assessed S&P skills. Still, the gender effect persists even if we control for IT experience. The regression results showed that women are 47% less likely to identify themselves as experts compared to men; while a participant who has no IT experience is 71% less likely to identify themselves as expert compared to a participant who has IT experience (both effects are significant with p<.001).

Summary.   Men indicated greater levels of security attitude on the scales engagement, attentiveness, and resistance, and greater technological affinity. In comparison, women indicated greater levels of security attitude on the scale concernedness, and greater privacy concerns. Further, men and women were interested in preventing different S&P risk scenarios. There were no gender differences in terms of security behavioral intention. In addition, men performed better in terms of technical privacy tools knowledge and technical privacy literacy, and indicated a greater familiarity, self-confidence, knowledge, and skills with regards to S&P than women.

5   Discussion

We expected the findings of our study in the S&P field to mirror those previously reported for mathematics [16], i.e., that stereotype-laden videos can elicit the stereotype threat effect and consequently influence performance and S&P attitudes. When interpreting our findings, it is important to to consider that there is an essential difference between the prior studies on mathematics and our study: while the mathematics studies were able to objectively measure performance in mathematics tests, our study relies on self-reported data.

Based on these subjective self-reports, we could not find broad evidence for a stereotype threat effect from advertisement and campaign videos in the security and privacy field. The first effect, namely that participating women in the *Non-stereotype* condition reported more interest to prevent hacker access than those in the stereotype conditions, is in line with expectations and might indicate that stereotype threat effects can occur. In contrast, the second effect, namely that participating women in the *Stereotype women* condition reported higher levels of computer security knowledge than participating women in the *Stereotype men* condition, does not seem to relate to the stereotype threat effect. The effect seems to be rather due to problems with eliciting the priming (as discussed

in the next section), due to an anomaly in our sampling that lead to this effect, or due to different factors that influenced our participants' gender attitudes across their lifetime.

Several such factors could have played a role and over-shadowed the priming in our study. Firstly, if participating women are affected by an unwelcoming or unsupported environment and no support infrastructure is in place to counter-act the environmental influences, that might have detrimental effects [11, 57]. Secondly, gender norms might have been adopted by participants due to interpersonal influences in their early adolescence which might in turn have perpetuated stereotypical attitudes far deeper than our study priming could [38]. Concrete results of these differences could be that it is seen more acceptable to not be knowledgeable in the security and privacy field or that different sources for information about security and privacy topics are considered [12]. In any case, further investigations are needed in order to gather further evidence relating to these effects.

Priming Studies.   While the effect we found in our data (increased interest in preventing hackers from getting access to devices in the *Non-stereotype* condition) might be an artifact stemming from the sample, it is also possible that the technical priming from the video clips in the *Non-stereotype* condition made the concepts of hacking and device protection more prevalent in the participating women's minds. Participating men who were shown the same video clips did not report increased interest in preventing hacker access to their devices; still, participating men might identify less with the protagonists in the video clips, who are girls and women. Priming studies generally present a number of challenges: For example, the duration of the priming should be chosen with care [71], participants may react with reactance to being influenced [17], especially if the stimuli contain such a clear message as in the *Non-stereotype* condition videos from the #SheCanSTEM and Dare to STEM campaigns. Further, if the content is perceived to be unrealistic, participants' response to a stimulus might be delayed [35]. Also, our priming to-wards S&P gender stereotypes was rather subtle, as our study did not include a condition with videos showing women performing poorly at STEM tasks. Even using the wide range of search terms described in Section 3.1, we could not find any commercials that fell into this category. Therefore, we could not include this as study condition and had to rely on the *Stereotype men* condition instead which portrayed men but not women as proficient in STEM. The *Stereotype women* condition, on the other hand, depicted women as mothers or spouses. Hence, participating women who are not mothers or in a relationship may not have identified with the women portrayed in the videos. To aid researchers in selecting appropriate priming stimuli, we advocate the creation of databases with validated gender priming content, as has been done in other research domains (e.g., [34]). Finally, priming studies might fail to overcome stereotypes which have been engrained

from early childhood on with one-shot stimuli exposure. In the following, we thus discuss alternative paths to overcome gender stereotypes in the S&P context.

Acknowledge the Gender-Imbalance in Today's Ads. The videos we used in our study were not the most recent ads by the respective companies and actually up to 10 years old. This raises the question of whether current advertisement videos are less prone to depict gender stereotypes and still represent suitable objects of study for our experiment. After all, if newer ads do not rely on gender stereotyping, investigating other media might have been the more prudent way to go. However, from analyses of the literature, we know that this is not the case [49]: Women are still depicted as caregivers and men as more independent. This manifests in the continuous need to review and ban advertisements for inappropriate portrayals by authorities, as has been done recently, e.g., for Aptamil in the UK[12] or Honey Birdette in Australia[13]. Thus, the gender imbalance is still there, even with the twist that men are increasingly sexualized and objectified as well (though substantially less than women) [32, 40]. While we explicitly decided against using such banned ads, our results indicate that stereotype threat is currently less of a concern and the issue might lie deeper entrenched in the social norms of societies and the cognitive maps of the children in these societies.

Address Self-Concepts of Children. According to Gottfredson [26], young people start to develop self-concepts that shape their cognitive map of preferences, interests, and aspired competencies early on. Gottfredson [26] and Erdmann et al. [19] posit that gender is one of the most salient cues for selecting role models that serve as direction for these self-concepts, thus, young people mostly lean towards gender-typical options. As a result, gender-atypical options that are not part of their cognitive maps might not even be on a person's radar as they get older. Based on these considerations, Erdmann et al. [19] advocate long-term counseling for young people to break up stereotypical educational choices above short-term interventions that provide too few new experiences to alter a person's cognitive map. They further assume that people can only become role models if they have a close relationship with the addressee, which is usually not the case in short-term interventions.

Hence, it is possible that the campaign videos in our study, being a prime example for short-term interventions, had no significant effect on our participants since they failed to modify the participants' cognitive maps and to provide adequate role models. Long-term interventions such as counseling [19]

---

or mentoring could thus be promising measures for countering stereotypes by pointing out gender-atypical options and support recipients in sticking to their choice, even if this means violating social norms. Campaign videos such as the clips used in our study could then be launched to advertise such long-term programs.

In addition, campaigns should embrace existing role models with whom the addressees already identify. For this, it might also be feasible to inspire communication about S&P topics between less experienced users and people from their social environment who are proficient in these topics, and who could then become role models [22, 43, 56]. Likewise, S&P advocates who serve as role models in a professional context [29, 30, 31, 64] could broach existing stereotypes directly to sensitize their audience to this issue. Still, given the already existing under-representation of women in the S&P field, care has to be taken as not to place additional burdens on those women and thereby intensify unequal job conditions. This could possibly be addressed, e.g., by offering mentoring or orientation programs in which experts from different gender groups participate.

Adopt S&P Content in Curricula. One striking difference between the S&P field and mathematics is that math is a mandatory school subject and so everyone who has undergone the same level of schooling is exposed to roughly the same material (even if some take away more from lessons than others). In multiple countries, this is not the case for S&P subjects, where the acquisition of knowledge and skills is largely dependent on a pre-existing interest in the matter and must be done in one's free time (excluding mandatory workplace S&P programs of questionable quality which set in much later than schooling). Such interests might themselves be driven by social norms and stereotypes [9, 62] and thus any stereotype threat effect might (on average) be overshadowed by actual differences in knowledge and interest resulting from these social influences. Since women have been also found to underestimate their competencies compared to men [15], repeated positive experiences might strengthen girls' and women's self-efficacy. Thus, promising avenues to accustom young people to technical and S&P content regardless of contradicting self-concepts and gender norms include, e.g., integrating such content in existing mandatory school subjects, as mandatory content in training for non-technical professions, or as applications in suitable non-technical degree programs such as economics, social sciences, and law.

Improve Gender Representations at Large Scales. An alternative approach that goes deeper to the root of the problem would be to avoid exposing children to social gender norms and stereotypes. This would require, for example, to transition towards a market with more gender-neutral toys, advertising messages, and content in fictional and factual media such as school textbooks. In addition, gender stereotypes already

---

established in society at large would have to be addressed directly by parents and teachers and exposed as such. To this end, it might be beneficial to address awareness campaigns and counseling programs not only directly to young people or those affected by stereotypes, but also specifically to parents and teachers. A first step towards this is the "#EndGender-Stereotypes" campaign launched by the European Commission in 2023 [68], which aims to challenge widespread gender stereotypes and targets the entire society.

In the U.S., the STEM Opportunities Act [33] seeks to clear the path for people from groups that have been historically underrepresented in the STEM fields, including women, to pursue careers in STEM. Measures include, for example, the organization of workshops that raise awareness for this issue at universities and federal science agencies, and the funding of research work on this topic. The goal thereby should be to have offers available also for marginalized communities and remote areas to reach individuals that might otherwise be excluded from such opportunities.

Future Work. Considering our results, in particular, longitudinal studies investigating when S&P-related social norms are formed seem to be an important line of future work. By that, we echo other work [38]. Such research would shed light on the mechanisms that underlie the prevalent gender imbalances in the S&P field. Based on such studies it would be possible to inform the development and recommendation of interventions tailored to the age when they are most relevant, e.g., campaigns focused on certain age brackets including materials and information for parents that want to prevent such social norms from manifesting in their children's self-concepts.

Additionally, it might be worthwhile to investigate whether it was our method of elicitation that did not have the intended effect (despite our manipulation check). Using different methodologies, e.g., based on direct communication to counter the stereotypes [37], might yield different results, albeit we are skeptical of this.

Ideally, future studies would complement self-reported data with objective metrics. Objective metrics might include knowledge questions as used in testing the effectiveness of security and privacy awareness or education materials (e.g., [45]) or as used in other human factors studies investigating behavior (e.g., [69]). However, in selecting such tasks, care should be taken to not introduce different bias into the study design. For instance multiple different domains in security and privacy would need to be covered and comparable difficulty should be ensured.

6 Conclusion

Inspired by similar research in the field of mathematics [16], we conducted an experimental between-subject study with 959 participants recruited via Prolific to explore whether videos (1) depicting stereotypes associated with women, (2) stereotypes associated with men, (3) non-stereotype depictions, and (4) a control condition showing only non-anthropomorphic content influence women and men in the security and privacy (S&P) field. We find few effects of the videos, but our results show that women who had been exposed to non-stereotype videos reported more interest in preventing hacker access to their devices. In addition, our findings indicate a variety of gender differences, with men reporting higher levels of S&P intentions, and knowledge, while women report higher levels of S&P concern. Based on our findings, we derive several implications for addressing gender stereotypes and social norms, such as implementing long-term interventions (e.g., counseling or mentoring) that target children, young adults, but also parents and teachers, emphasizing familiar people as S&P role models, and exposing students to gender-atypical content via S&P curricula.

Acknowledgments

References

[1] Desiree Abrokwa, Shruti Das, Omer Akgul, and Michelle L. Mazurek. Comparing security and privacy attitudes among U.S. users of different smartphone and Smart-Speaker platforms. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 139–158. USENIX Association, August 2021. Available at https://www.usenix.org/conference/soups2021/presentation/abrokwa.

[2] Adam L. Alter, Joshua Aronson, John M. Darley, Cordaro Rodriguez, and Diane N. Ruble. Rising to the threat: Reducing stereotype threat by reframing the threat as a challenge. *Journal of Experimental Social Psychology*, 46(1):166–171, January 2010. doi: 10.1016/j.jesp.2009.09.014.

[3] Joshua Aronson, Carrie B. Fried, and Catherine Good. Reducing the Effects of Stereotype Threat on African American College Students by Shaping

Theories of Intelligence. *Journal of Experimental Social Psychology*, 38(2):113–125, March 2002. doi: 10.1006/jesp.2001.1491.

[4] Carol J. Auster and Claire S. Mansbach. The gender marketing of toys: An analysis of color and type of toy on the disney store website. *Sex roles*, 67:375–388, 2012. doi: 10.1007/s11199-012-0177-8.

[5] Davide Barbieri, Jakub Caisl, Marre Karu, Giulia Lanfredi, Blandine Mollard, Vytautas Peciukonis, Maria Belen Pilares La Hoz, Jolanta Reingardė, and Lina Salanauskaitė. Gender Equality Index 2020 - Digitalisation and the future of work, 2020. Available at https://eige.europa.eu/publications-resources/publications/gender-equality-index-2020-digitalisation-and-future-work (Accessed 07-May-2024).

[6] Carlos Bermejo Fernandez, Dimitris Chatzopoulos, Dimitrios Papadopoulos, and Pan Hui. This website uses nudging: Mturk workers' behaviour on cookie consent notices. *Proceedings of the ACM on human-computer interaction*, 5(CSCW2), October 2021. doi: 10.1145/3476087.

[7] Bianca Bush and Adrian Furnham. Gender jenga: the role of advertising in gender stereotypes within educational and non-educational games. *Young Consumers*, 14(3):216–229, 2013. doi: 10.1108/YC-11-2012-00324.

[8] James Carifio and Rocco Perla. Ten common misunderstandings, misconceptions, persistent myths and urban legends about likert scales and likert response formats and their antidotes. *Journal of Social Sciences*, 3(3), 03 2007. doi: 10.3844/jssp.2007.106.116.

[9] Sapna Cheryan, Victoria C. Plaut, Caitlin Handron, and Lauren Hudson. The Stereotypical Computer Scientist: Gendered Media Representations as a Barrier to Inclusion for Women. *Sex Roles*, 69(1-2):58–71, July 2013. doi: 10.1007/s11199-013-0296-x.

[10] Jacob Cohen. A power primer. *Psychological Bulletin*, 112(1):155–159, 07 1992. doi: 10.1037/0033-2909.112.1.155.

[11] J. McGrath Cohoon. Recruiting and retaining women in undergraduate computing majors. *ACM SIGCSE Bulletin*, 34(2):48–52, June 2002. doi: 10.1145/543812.543829.

[12] Kovila P. L. Coopamootoo and Magdalene Ng. "Un-Equal online safety?" a gender analysis of security and privacy protection advice and behaviour patterns. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5611–5628, Anaheim, CA, August 2023. USENIX Association. Available at https://www.usenix.org/conference/usenixsecurity23/presentation/coopamootoo.

[13] Kovila P.L. Coopamootoo, Maryam Mehrnezhad, and Ehsan Toreini. "i feel invaded, annoyed, anxious and i may protect myself": Individuals' feelings about online tracking and their protective behaviour across gender and country. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 287–304, Boston, MA, August 2022. USENIX Association. Available at https://www.usenix.org/conference/usenixsecurity22/presentation/coopamootoo.

[14] Sheila J. Cunningham and C. Neil Macrae. The colour of gender stereotyping. *British Journal of Psychology*, 102(3):598–614, 2011. doi: 10.1111/j.2044-8295.2011.02023.x.

[15] Initiative D21. D21 digital gender gap, 2020. Accessed from https://initiatived21.de/uploads/03_Studien-Publikationen/Digital-Gender-Gap/d21_digitalgendergap.pdf (Accessed 07-May-2024).

[16] Paul G. Davies, Steven J. Spencer, Diane M. Quinn, and Rebecca Gerhardstein. Consuming Images: How Television Commercials that Elicit Stereotype Threat Can Restrain Women Academically and Professionally. *Personality and Social Psychology Bulletin*, 28(12):1615–1628, December 2002. doi: 10.1177/014616702237644.

[17] Kai Duttle and Tatsuhiro Shichijo. Default or reactance? Identity priming effects on overconfidence in Germany and Japan. Working Papers on East Asian Studies 103/2015, University of Duisburg-Essen, Institute of East Asian Studies IN-EAST, 2015. Available at https://ideas.repec.org/p/zbw/udedao/1032015.html.

[18] Serge Egelman and Eyal Peer. Scaling the security wall: Developing a security behavior intentions scale (SeBIS). In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 2873—-2882, New York, NY, USA, 2015. Association for Computing Machinery. doi: 10.1145/2702123.2702249.

[19] Melinda Erdmann, Juliana Schneider, Irena Pietrzyk, Marita Jacob, and Marcel Helbig. The impact of guidance counselling on gender segregation: Major choice and persistence in higher education. An experimental study. *Frontiers in Sociology*, 8:1154138, April 2023. doi: 10.3389/fsoc.2023.1154138.

[20] Cori Faklaris, Laura Dabbish, and Jason I. Hong. Do they accept or resist cybersecurity measures? development and validation of the 13-item security attitude inventory (sa-13), 2022.

[21] Thomas Franke, Christiane Attig, and Daniel Wessel. A personal resource for technology interaction: Development and validation of the affinity for technology interaction (ATI) scale. *International Journal of Human–Computer Interaction*, 35(6):456–467, 2019. doi: 10.1080/10447318.2018.1456150.

[22] Nina Gerber and Karola Marky. The nerd factor: The potential of {S&P} adepts to serve as a social resource in the user's quest for more secure and {Privacy-Preserving} behavior. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 57–76, 2022.

[23] Peter Glick, Susan Fiske, Antonio Mladinic, José Saiz, Dominic Abrams, Barbara Masser, Bolanle Adetoun, Johnston Osagie, Adebowale Akande, Amos Alao, Annetje Brunner, Tineke Willemsen, Kettie Chipeta, Benoit Dardenne, Ap Dijksterhuis, Daniel Wigboldus, Thomas Eckes, Iris Six-Materna, Francisca Expósito, and Wilson López-López. Beyond prejudice as simple antipathy: Hostile and benevolent sexism across cultures. *Journal of personality and social psychology*, 79(5):763–75, 11 2000. doi: 10.1037/0022-3514.79.5.763.

[24] Antonio Gómez-Aguilar. Content bubbles: How platforms filter what we see. In *Handbook of Research on Transmedia Storytelling, Audience Engagement, and Business Strategies*, pages 338–350. IGI Global, 2020.

[25] Catherine Good, Joshua Aronson, and Michael Inzlicht. Improving adolescents' standardized test performance: An intervention to reduce the effects of stereotype threat. *Journal of Applied Developmental Psychology*, 24(6):645–662, December 2003. doi: 10.1016/j.appdev.2003.09.002.

[26] Linda S Gottfredson. Circumscription and compromise: A developmental theory of occupational aspirations. *Journal of Counseling psychology*, 28(6):545, 1981.

[27] Thomas Gross. Validity and reliability of the scale internet users' information privacy concerns (IUIPC). *Proceedings on Privacy Enhancing Technologies*, 2021: 235 – 258, 2021.

[28] Lauren Gurrieri, Mandy McKenzie, and Megan Bugden. Community Responses to Gender Portrayals in Advertising. *Women's Health Victoria Issue Papers*, (15), December 2019. Available at https://shequal.com.au/app/uploads/2023/05/Issues-Paper_2019.10.29_Community-responses-to-gender-portrayals-in-advertising-a-research-paper_Fulltext-PDF.pdf.

[29] Julie Haney and Wayne Lutters. Cybersecurity advocates: Discovering the characteristics and skills for an emergent role. *Information & Computer Security*, 29, 2021.

[30] Julie Haney, Wayne Lutters, and Jody Jacobs. Cybersecurity advocates: Force multipliers in security behavior change. *IEEE Security & Privacy*, 19(4): 54–59, 2021. doi: 10.1109/MSEC.2021.3077405.

[31] Julie M. Haney and Wayne G. Lutters. "it's Scary...It's Confusing...It's dull": How cybersecurity advocates overcome negative perceptions of security. In *Fourteenth Symposium on Usable Privacy and Security*, SOUPS 2018, pages 411–425, Baltimore, MD, August 2018. USENIX Association. Available at https://www.usenix.org/conference/soups2018/presentation/haney-perceptions.

[32] Erin Hatton and Mary Nell Trautner. Equal opportunity objectification? the sexualization of men and women on the cover of rolling stone. *Sexuality & culture*, 15: 256–278, 2011.

[33] H.R.204–117th Congress. STEM opportunities act, 2021. Available at https://www.congress.gov/bill/117th-congress/house-bill/204 (Accessed 15-February-2024).

[34] Keith A. Hutchison, David A. Balota, James H. Neely, Michael J. Cortese, Emily R. Cohen-Shikora, Chi-Shing Tse, Melvin J. Yap, Jesse J. Bengson, Dale Niemeyer, and Erin Buchanan. The semantic priming project. *Behavior Research Methods*, 45(4):1099–1114, December 2013. doi: 10.3758/s13428-012-0304-z.

[35] Jakob D. Jensen, Jennifer K. Bernat, Kari M. Wilson, and Julie Goonewardene. The delay hypothesis: The manifestation of media effects over time. *Human Communication Research*, 37(4):509–528, 2011. doi: 10.1111/j.1468-2958.2011.01415.x.

[36] Johnny V. Sparks John G. Wirtz and Thais M. Zimbres. The effect of exposure to sexual appeals in advertisements on memory, attitude, and purchase intention: A meta-analytic review. *International Journal of Advertising*, 37(2):168–198, 2018. doi: 10.1080/02650487.2017.1334996.

[37] Michael Johns, Toni Schmader, and Andy Martens. Knowing Is Half the Battle: Teaching Stereotype Threat as a Means of Improving Women's Math Performance. *Psychological Science*, 16(3):175–179, March 2005. doi: 10.1111/j.0956-7976.2005.00799.x.

[38] Anna Kågesten, Susannah Gibbs, Robert Wm Blum, Caroline Moreau, Venkatraman Chandra-Mouli, Ann Herbert, and Avni Amin. Understanding factors that shape gender attitudes in early adolescence globally: A mixed-methods systematic review. *PloS one*, 11(6): e0157805, 2016. doi: 10.1371/journal.pone.0157805.

[39] Ruogu Kang, Laura Dabbish, Nathaniel Fruchter, and Sara Kiesler. "My data just goes Everywhere:" user mental models of the internet and implications for privacy and security. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 39–52, Ottawa, July 2015. USENIX Association. Available at https://www.usenix.org/conference/soups2015/proceedings/presentation/kang.

[40] Kathrin Karsay, Johannes Knoll, and Jörg Matthes. Sexualizing media use and self-objectification: A meta-analysis. *Psychology of women quarterly*, 42(1): 9–28, 2018.

[41] Agnieszka Kitkowska, Farzaneh Karegar, and Erik Wästlund. Share or protect: Understanding the interplay of trust, privacy concerns, and data sharing purposes in health and well-being apps. In *Proceedings of the 15th Biannual Conference of the Italian SIGCHI Chapter*, pages 1–14, 2023.

[42] Jenny Lewin-Jones and Barbara Mitra. Gender roles in television commercials and primary school children in the uk. *Journal of children and media*, 3(1):35–50, 2009.

[43] Heather Richter Lipford and Mary Ellen Zurko. Someone to watch over me. In *Proceedings of the 2012 New Security Paradigms Workshop*, NSPW '12, pages 67––76, New York, NY, USA, 2012. Association for Computing Machinery. doi: 10.1145/2413296.2413303.

[44] Naresh K. Malhotra, Sung S. Kim, and James Agarwal. Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *Information Systems Research*, 15(4):336–355, 2004. doi: 10.1287/isre.1040.0032.

[45] Peter Mayer, Christian Schwartz, and Melanie Volkamer. On The Systematic Development and Evaluation Of Password Security Awareness-Raising Materials. Annual Computer Security Applications Conference, pages 733 – 748, 2018. doi: 10.1145/3274694.3274747.

[46] Peter Mayer, Yixin Zou, Florian Schaub, and Adam J. Aviv. "now i'm a bit angry:" individuals' awareness, perception, and responses to data breaches that affected them. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 393–410. USENIX Association, August 2021. Available at https://www.usenix.org/conference/usenixsecurity21/presentation/mayer.

[47] Matthew S. McGlone and Joshua Aronson. Stereotype threat, identity salience, and spatial reasoning. *Journal of Applied Developmental Psychology*, 27(5):486–493, September 2006. doi: 10.1016/j.appdev.2006.06.003.

[48] Rusty B. McIntyre. Alleviating women's mathematics stereotype threat through salience of group achievements. *Journal of Experimental Social Psychology*, 2003.

[49] Mandy McKenzie, Megan Bugden, Webster Amy, and Mischa Barr. Advertising (In)Equality - The Impacts of Sexist Advertising on Women's Health and Wellbeing. *Women's Health Issues Paper*, (14), December 2018.

[50] Joel T. Nadler and M. H. Clark. Stereotype threat: A meta-analysis comparing african americans to hispanic americans. *Journal of Applied Social Psychology*, 41(4):872–890, 2011. doi: 10.1111/j.1559-1816.2011.00739.x.

[51] N. Narasimhamurthy. Television advertisement and its impact on attitudes, behaviors of children-a study. *International Journal of interdisciplinary and multidisciplinary Studies*, 1(10):14–22, 2014.

[52] Charlie Osborne. Women to hold 30 percent of cybersecurity jobs globally by 2025, 2023. Available at https://cybersecurityventures.com/women-in-cybersecurity-report-2023/ (Accessed 07-May-2024).

[53] Adam Peruta and Jack Powers. Look who's talking to our kids: Representations of race and gender in tv commercials on nickelodeon. *International Journal of Communication*, 11:16, 2017.

[54] Katherine Picho and Scott W. Brown. Can Stereotype Threat Be Measured? A Validation of the Social Identities and Attitudes Scale (SIAS). *Journal of Advanced Academics*, 22(3):374–411, May 2011. doi: 10.1177/1932202X1102200302.

[55] Jennifer J. Pike and Nancy A. Jennings. The effects of commercials on children's perceptions of gender appropriate toy use. *Sex roles*, 52:83–91, 2005.

[56] Elissa M. Redmiles, Sean Kross, and Michelle L. Mazurek. How i learned to be secure: A census-representative survey of security advice sources and behavior. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, pages 666––677, New York, NY, USA,

2016. Association for Computing Machinery. doi: 10.1145/2976749.2978307.

[57] Penny Rheingans, Erica D'Eramo, Crystal Diaz-Espinoza, and Danyelle Ireland. A Model for Increasing Gender Diversity in Technology. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, pages 459–464, Baltimore Maryland USA, February 2018. ACM. doi: 10.1145/3159450.3159533.

[58] Yukiko Sawaya, Mahmood Sharif, Nicolas Christin, Ayumu Kubota, Akihiro Nakarai, and Akira Yamada. Self-confidence trumps knowledge: A cross-cultural study of security behavior. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 2202–2214, New York, NY, USA, 2017. Association for Computing Machinery. doi: 10.1145/3025453.3025926.

[59] Steve Sheng, Mandy Holbrook, Ponnurangam Kumaraguru, Lorrie Faith Cranor, and Julie Downs. Who falls for phish? a demographic analysis of phishing susceptibility and effectiveness of interventions. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 373–382, 2010.

[60] Arielle M. Silverman and Geoffrey L. Cohen. Stereotypes as stumbling-blocks: How coping with stereotype threat affects life outcomes for people with physical disabilities. *Personality and Social Psychology Bulletin*, 40(10):1330–1340, 2014. doi: 10.1177/0146167214542800.

[61] Steven J. Spencer, Claude M. Steele, and Diane M. Quinn. Stereotype Threat and Women's Math Performance. *Journal of Experimental Social Psychology*, 35 (1):4–28, January 1999. doi: 10.1006/jesp.1998.1373.

[62] Christine R. Starr. "I'm Not a Science Nerd!": STEM Stereotypes, Identity, and Motivation Among Undergraduate Women. *Psychology of Women Quarterly*, 42(4):489–503, December 2018. doi: 10.1177/0361684318793848.

[63] Peter Story, Daniel Smullen, Yaxing Yao, Alessandro Acquisti, Lorrie Cranor, Norman Sadeh, and Florian Schaub. Awareness, adoption, and misconceptions of web privacy tools. *Proceedings on Privacy Enhancing Technologies*, 2021(3):308–333, 07 2021. doi: 10.2478/popets-2021-0049.

[64] Mohammad Tahaei, Alisa Frik, and Kami Vaniea. Privacy champions in software teams: Understanding their motivations, strategies, and challenges. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2021. Association for Computing Machinery. doi: 10.1145/3411764.3445768.

[65] Jenny Tang, Eleanor Birrell, and Ada Lerner. Replication: How well do my results generalize now? the external validity of online privacy and security surveys. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 367–385, Boston, MA, August 2022. USENIX Association. Available at https://www.usenix.org/conference/soups2022/presentation/tang.

[66] Theresa Tran and Hilary Yerbury. New perspectives on personalised search results: Expertise and institutionalisation. *Australian Academic & Research Libraries*, 46(4):277–290, 2015. doi: 10.1080/00048623.2015.1077302.

[67] Sabine Trepte, Doris Teutsch, Philipp K. Masur, Carolin Eicher, Mona Fischer, Alisa Hennhöfer, and Fabienne Lind. *Do People Know About Privacy and Data Protection Strategies? Towards the "Online Privacy Literacy Scale" (OPLIS)*, pages 333–365. Springer Netherlands, Dordrecht, 2015. doi: 10.1007/978-94-017-9385-8_14.

[68] European Union. #endgenderstereotypes, 2023. Available at https://end-gender-stereotypes.campaign.europa.eu/ (Accessed 15-February-2024).

[69] Blase Ur, Jonathan Bees, Sean M Segreti, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. Do Users' Perceptions of Password Security Match Reality? In *Conference on Human Factors in Computing Systems*, Conference on Human Factors in Computing Systems, pages 3748 – 3760, 2016. doi: 10.1145/2858036.2858546.

[70] Miranda Wei, Pardis Emami-Naeini, Franziska Roesner, and Tadayoshi Kohno. Skilled or Gullible? Gender Stereotypes Related to Computer Security and Privacy. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 2050–2067. IEEE. doi: 10.1109/SP46215.2023.10179469.

[71] Dirk Wentura and Klaus Rothermund. Priming is not priming is not priming. *Social Cognition*, 32 (Supplement):47–67, 2014.

Appendix

Table 2: ANOVA results comparing security attitude (SA-13 [20]), technological affinity (ATI [21]), privacy concerns (IUIPC-8 [27, 44]), and security behavior intention (SeBIS [18]) (DV) between the different priming groups (IV) for women. In cases where homogeneity of variance was not given, Welch's ANOVA was calculated.

	df	F-value	Sig.	Control		Stereotype men		Stereotype women		Non-stereotype	
				M	SD	M	SD	M	SD	M	SD
SA13_Engagement	3, 461	0.304	.823	3.24	0.95	3.23	1.01	3.20	0.98	3.32	0.92
SA13_Attentiveness	3, 461	1.413	.238	3.29	0.83	3.32	0.96	3.25	0.81	3.47	0.87
SA13_Resistance	3, 461	0.793	.498	2.38	0.81	2.32	0.80	2.35	0.82	2.22	0.77
SA13_Concernedness	3, 461	2.169	.091	3.78	0.82	3.75	0.87	3.55	0.91	3.58	0.81
SEBIS_DeviceSecurement	3, 461	0.445	.721	3.92	0.88	3.83	0.98	3.92	0.92	3.97	0.94
SEBIS_Updating	3, 461	1.570	.196	3.66	0.96	3.63	0.94	3.43	0.90	3.63	0.83
SEBIS_PasswordGeneration	3, 461	0.787	.501	3.73	0.80	3.65	0.86	3.58	0.95	3.72	0.85
SEBIS_ProactiveAwareness	3, 255.302	0.824	.481	3.83	0.68	3.77	0.75	3.73	0.87	3.87	0.76
ATI_Overall	3, 461	0.655	.580	3.34	1.04	3.25	1.19	3.45	1.05	3.33	1.18
IUIPC_Control	3, 461	0.915	.434	6.09	1.13	6.30	0.91	6.12	1.09	6.20	0.97
IUIPC_Awareness	3, 461	0.320	.811	6.53	0.89	6.57	0.77	6.49	0.86	6.58	0.80
IUIPC_Collection	3, 461	2.018	.111	6.04	1.17	5.99	1.20	5.79	1.17	6.14	0.95

Table 3: ANOVA results comparing security attitude (SA-13 [20]), technological affinity (ATI [21]), privacy concerns (IUIPC-8 [27, 44]), and security behavior intention (SeBIS [18]) (DV) between the different priming groups (IV) for men. In cases where homogeneity of variance was not given, Welch's ANOVA was calculated.

	df	F-value	Sig.	Control		Stereotype men		Stereotype women		Non-stereotype	
				M	SD	M	SD	M	SD	M	SD
SA13_Engagement	3, 257.066	0.228	.877	3.44	0.91	3.45	1.00	3.49	0.83	3.40	0.83
SA13_Attentiveness	3, 256.540	0.774	.510	3.48	0.84	3.51	0.94	3.60	0.71	3.46	0.78
SA13_Resistance	3, 578	0.695	.555	2.40	0.71	2.51	0.84	2.39	0.74	2.38	0.84
SA13_Concernedness	3, 578	0.162	.922	3.37	0.94	3.29	0.89	3.33	0.85	3.35	0.83
SEBIS_DeviceSecurement	3, 578	0.665	.574	3.93	0.96	3.98	0.96	3.84	0.83	3.99	0.90
SEBIS_Updating	3, 578	0.306	.821	3.75	0.86	3.69	0.88	3.64	0.84	3.70	0.78
SEBIS_PasswordGeneration	3, 256.828	0.368	.776	3.70	0.86	3.66	0.91	3.59	0.81	3.68	0.72
SEBIS_ProactiveAwareness	3, 578	0.241	.868	3.74	0.79	3.68	0.81	3.76	0.69	3.70	0.78
ATI_Overall	3, 578	2.549	.055	3.86	1.09	3.75	1.08	4.10	0.96	3.97	0.94
IUIPC_Control	3, 578	0.468	.705	5.87	1.21	5.74	1.24	5.83	1.09	5.91	1.13
IUIPC_Awareness	3, 578	0.245	.865	6.26	1.08	6.16	1.16	6.28	1.07	6.23	1.14
IUIPC_Collection	3, 578	0.539	.656	5.71	1.32	5.50	1.39	5.55	1.23	5.60	1.27

Table 4: Kruskal-Wallis test results comparing interest to prevent various S&P risk scenarios (DV) between the different priming groups (IV) for women.

	df	H-value	Sig.	Control M_{rank}	Stereotype men M_{rank}	Stereotype women M_{rank}	Non-stereotype M_{rank}
Prevent: hackers from gaining access to your device	3	11.696	.009**	238.86	220.41	220.99	251.95
Prevent: online stores from misusing your credit card information	3	7.227	.065	235.92	219.28	227.79	249.09
Prevent: advertisers from seeing the website you visit	3	1.679	.642	231.49	236.04	222.05	242.73
Prevent: advertisers from showing you targeted ads based on the websites you visit	3	3.582	.310	227.89	236.57	218.69	249.31
Prevent: the websites you visit from seeing what physical location you are browsing from	3	3.851	.278	244.36	215.44	231.24	240.80
Prevent: your search engine from personalizing the search results you see based on the websites you visit	3	0.183	.980	236.47	229.28	232.65	233.55
Prevent: your internet service provider from seeing the websites you visit	3	1.162	.762	227.32	236.43	226.78	241.73
Prevent: the government from seeing the websites you visit	3	1.411	.703	223.99	230.73	234.32	243.08
Prevent: friends or family with physical access to your device from seeing the websites you visit in your browser history	3	4.113	.250	218.50	223.50	243.68	246.29
Prevent: your employer from seeing the websites you visit on your personal device while connected to your work's WiFi	3	1.796	.616	235.20	222.81	244.08	229.58
Prevent: law enforcement from seeing the websites you visit	3	7.383	.061	209.03	232.13	254.53	236.17
Prevent: companies who own movies from seeing if you illegally stream a movie	3	2.117	.548	222.76	226.03	240.28	242.91

*Note: *p<.05, **p<.01, ***p<.001*

Table 5: Kruskal-Wallis test results comparing interest to prevent various S&P risk scenarios (DV) between the different priming groups (IV) for men.

	df	H-value	Sig.	Control M$_{rank}$	Stereotype men M$_{rank}$	Stereotype women M$_{rank}$	Non-stereotype M$_{rank}$
Prevent: hackers from gaining access to your device	3	0.755	.860	241.23	231.88	231.47	233.33
Prevent: online stores from misusing your credit card information	3	1.153	.764	228.75	238.34	240.65	230.34
Prevent: advertisers from seeing the website you visit	3	1.250	.741	243.86	225.11	234.91	233.95
Prevent: advertisers from showing you targeted ads based on the websites you visit	3	6.139	.105	257.91	216.67	231.42	231.65
Prevent: the websites you visit from seeing what physical location you are browsing from	3	0.488	.922	231.34	230.52	240.87	235.26
Prevent: your search engine from personalizing the search results you see based on the websites you visit	3	2.495	.476	250.50	230.91	225.80	230.62
Prevent: your internet service provider from seeing the websites you visit	3	0.217	.975	237.56	230.06	235.41	234.91
Prevent: the government from seeing the websites you visit	3	5.265	.153	249.56	214.46	231.23	242.45
Prevent: friends or family with physical access to your device from seeing the websites you visit in your browser history	3	2.529	.470	241.20	241.25	218.41	237.14
Prevent: your employer from seeing the websites you visit on your personal device while connected to your work's WiFi	3	2.826	.419	242.91	233.20	219.01	242.79
Prevent: law enforcement from seeing the websites you visit	3	3.843	.279	250.63	218.58	230.79	237.74
Prevent: companies who own movies from seeing if you illegally stream a movie	3	1.558	.669	230.15	240.22	243.09	224.63

Table 6: Kruskal-Wallis test results comparing knowledge test performance, and self-assessed knowledge and skills (DV) between the different priming groups (IV) for women.

	df	H-value	Sig.	Control M$_{rank}$	Stereotype men M$_{rank}$	Stereotype women M$_{rank}$	Non-stereotype M$_{rank}$
KnowledgePrivacyTools	3	0.649	.885	229.54	235.90	239.37	227.09
OPLIS_Technical	3	0.509	.917	239.88	231.96	232.00	228.07
KnowHowSelfReportScale	3	0.242	.971	231.03	230.03	232.97	238.01
SelfConfidenceSecurityKnowledge	3	0.138	.987	233.25	229.37	231.64	235.72
TechnicalKnowledge_ General	3	4.785	.188	233.03	221.90	254.29	222.22
TechnicalKnowledge_ ComputerSecurity	3	8.570	.036*	230.94	214.80	262.04	223.50
TechnicalKnowledge_ Privacy	3	4.808	.186	233.70	217.71	253.77	226.27
Skills	3	2.972	.396	235.17	216.95	240.90	238.73

*Note: *p<.05, **p<.01, ***p<.001*

Table 7: Kruskal-Wallis test results comparing knowledge test performance, and self-assessed knowledge and skills (DV) between the different priming groups (IV) for men.

	df	H-value	Sig.	Control M$_{rank}$	Stereotype men M$_{rank}$	Stereotype women M$_{rank}$	Non-stereotype M$_{rank}$
KnowledgePrivacyTools	3	4.947	.176	232.64	214.44	252.94	237.82
OPLIS_Technical	3	1.173	.760	227.60	240.45	228.99	241.07
KnowHowSelfReportScale	3	0.544	.909	229.47	239.53	230.41	238.68
SelfConfidenceSecurityKnowledge	3	1.935	.586	228.08	243.88	241.06	223.17
TechnicalKnowledge_ General	3	0.874	.832	235.28	235.49	241.50	225.74
TechnicalKnowledge_ ComputerSecurity	3	4.644	.200	223.21	243.66	251.00	220.32
TechnicalKnowledge_ Privacy	3	2.789	.425	218.66	239.82	245.94	233.75
Skills	3	0.080	.994	234.63	235.29	236.06	232.03

Table 8: Unpaired t-test results comparing security attitude (SA-13 [20]), technological affinity (ATI [21]), privacy concerns (IUIPC-8 [27, 44]), and security behavior intention (SeBIS [18]) (DV) between women and men (IV). In cases where homogeneity of variance was not given, Welch's t-test was calculated.

	df	t-value	Sig.	d	Women M	Women SD	Men M	Men SD
SA13_Engagement	931	-3.256	.001**	-0.213	3.25	0.96	3.45	0.89
SA13_Attentiveness	931	-3.289	.001**	-0.215	3.33	0.87	3.51	0.82
SA13_Resistance	931	-1.969	.049*	-0.129	2.32	0.80	2.42	0.78
SA13_Concernedness	931	5.795	<.001***	0.379	3.66	0.86	3.33	0.88
SEBIS_DeviceSecurement	931	-0.381	.703		3.91	0.93	3.93	0.91
SEBIS_Updating	923.605	-1.872	.061		3.59	0.91	3.69	0.84
SEBIS_PasswordGeneration	931	0.135	.892		3.67	0.87	3.66	0.83
SEBIS_ProactiveAwareness	931	1.620	.106		3.80	0.77	3.72	0.76
ATI_Overall	923.846	-8.211	<.001***	-0.538	3.34	1.11	3.92	1.03
IUIPC_Control	917.367	4.662	<.001***	0.305	6.18	1.03	5.84	1.17
IUIPC_Awareness	863.300	4.784	<.001***	0.313	6.54	0.83	6.23	1.11
IUIPC_Collection	914.466	5.009	<.001***	0.328	5.99	1.13	5.59	1.30

*Note: *p<.05, **p<.01, ***p<.001*

Table 9: Wilcoxon rank-sum test results comparing interest to prevent various S&P risk scenarios (DV) between women and men (IV).

	Z-value	Sig.	r	Women M_{rank}	Men M_{rank}
Prevent: hackers from gaining access to your device	-3.046	.002**	0.100	485.33	448.79
Prevent: online stores from misusing your credit card information	-2.992	.003**	0.098	485.80	448.32
Prevent: advertisers from seeing the website you visit	-2.730	.006**	0.089	489.81	444.33
Prevent: advertisers from showing you targeted ads based on the websites you visit	-1.345	.179	0.044	478.42	455.65
Prevent: the websites you visit from seeing what physical location you are browsing from	-2.930	.003**	0.096	491.11	443.04
Prevent: your search engine from personalizing the search results you see based on the websites you visit	-2.464	.014*	0.081	488.00	446.13
Prevent: your internet service provider from seeing the websites you visit	-0.121	.903	0.004	468.01	466.00
Prevent: the government from seeing the websites you visit	-0.946	.344	0.031	459.24	474.71
Prevent: friends or family with physical access to your device from seeing the websites you visit in your browser history	-3.231	.001**	0.106	439.63	494.19
Prevent: your employer from seeing the websites you visit on your personal device while connected to your work's WiFi	-1.575	.115	0.052	453.99	479.92
Prevent: law enforcement from seeing the websites you visit	-3.706	<.001***	0.121	435.65	498.15
Prevent: companies who own movies from seeing if you illegally stream a movie	-3.624	<.001***	0.119	436.17	497.63

*Note: *p<.05, **p<.01, ***p<.001*

Table 10: Wilcoxon rank-sum test results comparing knowledge test performance, and self-assessed knowledge and skills (DV) between women and men (IV).

	Z-value	Sig.	r	Women M_{rank}	Men M_{rank}
KnowledgePrivacyTools	-9.471	<.001***	0.310	384.43	549.04
OPLIS_Technical	-6.003	<.001***	0.197	416.50	517.18
KnowHowSelfReportScale	-9.039	<.001***	0.296	387.08	546.40
SelfConfidenceSecurityKnowledge	-6.741	<.001***	0.221	406.60	525.01
TechnicalKnowledge_ General	-9.009	<.001***	0.295	389.60	543.90
TechnicalKnowledge_ ComputerSecurity	-8.141	<.001***	0.267	396.52	537.02
TechnicalKnowledge_ Privacy	-8.183	<.001***	0.268	396.19	537.35
Skills	-6.525	<.001***	0.214	416.41	517.27

*Note: *p<.05, **p<.01, ***p<.001*

Table 11: Study participants' demographics per video priming group.

	Women								Men							
	Control		Stereotype men		Stereotype women		Non-stereotype		Control		Stereotype men		Stereotype women		Non-stereotype	
Age	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
18-20	/	/	/	/	5	4.2	1	0.9	2	1.7	2	1.7	2	1.7	4	3.4
21-25	9	7.7	13	11.3	11	9.3	11	9.6	7	5.9	18	15.5	15	12.8	5	4.3
26-30	12	10.3	15	13.0	14	11.9	13	11.3	31	26.3	20	17.2	21	17.9	15	12.8
31-35	16	13.7	18	15.7	17	14.4	15	13.0	19	16.1	20	17.2	19	16.2	23	19.7
36-40	20	17.1	9	7.8	13	11.0	11	9.6	18	15.3	15	12.9	19	16.2	19	16.2
41-45	12	10.3	6	5.2	14	11.9	10	8.7	12	10.2	7	6.0	18	15.4	8	6.8
46-50	8	6.8	11	9.6	16	13.6	12	10.4	11	9.3	9	7.8	9	7.7	15	12.8
51-55	8	6.8	14	12.2	6	5.1	12	10.4	7	5.9	10	8.6	3	2.6	8	6.8
56-60	13	11.1	10	8.7	9	7.6	12	10.4	3	2.5	5	4.3	8	6.8	11	9.4
61-65	8	6.8	8	7.0	2	1.7	11	9.6	4	3.4	6	5.2	2	1.7	5	4.3
66-70	8	6.8	5	4.3	10	8.5	5	4.3	3	2.5	3	2.6	1	0.9	3	2.6
71-75	2	1.7	4	3.5	1	0.8	1	0.9	1	0.8	/	/	/	/	1	0.9
76-80	1	0.9	2	1.7	/	/	1	0.9	/	/	/	/	/	/	/	
> 80	/	/	/	/	/	/	/	/	/	/	1	0.9	/	/	/	
Education																
School student	/	/	2	1.7	/	/	1	0.9	2	1.7	1	0.9	4	3.4	1	0.9
High School Diploma	42	35.9	31	27.0	46	39.0	40	34.8	44	37.3	35	30.2	36	30.8	39	33.3
Bachelor's Degree	49	41.9	53	46.1	44	37.3	41	35.7	44	37.3	54	46.6	51	43.6	46	39.3
Master's Degree	11	9.4	12	10.4	18	15.3	19	16.5	19	16.1	19	16.4	18	15.4	24	20.5
Ph.D. or higher	4	3.4	4	3.5	1	0.8	/	/	6	5.1	1	0.9	3	2.6	4	3.4
Other	11	9.4	13	11.3	8	6.8	13	11.3	3	2.5	6	5.2	5	4.3	3	2.6
Occupation																
Employed full time	43	36.8	41	35.7	59	50.0	45	39.1	70	59.3	66	56.9	74	63.2	71	60.7
Employed part-time	21	17.9	17	14.8	11	9.3	19	55.7	8	6.8	14	12.1	12	10.3	14	12.0
Unemployed and on the lookout	7	6.0	5	4.3	4	3.4	5	60.0	7	5.9	7	6.0	9	7.7	9	7.7
Unemployed and not on the lookout	2	1.7	1	0.9	3	2.5	/	/	2	1.7	1	0.9	2	1.7	2	1.7
Student	3	2.6	4	3.5	9	7.6	3	2.6	9	7.6	9	7.8	4	3.4	2	1.7
Retired	9	7.7	16	13.9	9	7.6	10	8.7	3	2.5	7	6.0	4	3.4	4	3.4
Homemaker	10	8.5	11	9.6	9	7.6	10	8.7	3	2.5	1	0.9	1	0.9	/	/
Self-employed	16	13.7	17	14.8	13	11.0	18	15.7	15	12.7	9	7.8	10	8.5	14	12.0
Incapacitated for work	3	2.6	2	1.7	1	0.8	5	4.3	1	0.8	/	/	1	0.9	1	0.9
Other	2	1.7	/	/	/	/	/	/	/	/	2	1.7	/	/	/	/
IT Experience																
Yes	23	19.7	24	20.9	21	17.8	21	18.3	49	41.5	50	43.1	50	42.7	53	45.3
No	95	81.2	91	79.1	96	81.4	94	81.7	68	57.6	64	55.2	66	56.4	61	52.1
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Hostile Sexism	2.768	0.766	2.820	0.788	2.689	0.808	2.696	0.721	3.054	0.892	3.083	0.895	3.070	0.841	3.076	0.879
Benevolent Sexism	3.056	0.768	3.187	0.862	3.061	0.841	3.172	0.749	3.214	0.869	3.161	0.844	3.262	0.766	3.365	0.689
Gender Identification	4.765	1.219	4.881	1.235	4.889	1.379	4.873	1.463	4.495	1.426	4.138	1.532	4.524	1.271	4.535	1.363

Towards Bridging the Research-Practice Gap:
Understanding Researcher-Practitioner Interactions and Challenges in Human-Centered Cybersecurity

Julie M. Haney, Clyburn Cunningham IV, and Susanne M. Furman
National Institute of Standards and Technology

Abstract

Human-centered cybersecurity (HCC) researchers seek to improve people's experiences with cybersecurity. However, a disconnect between researchers and practitioners, the *research-practice gap*, can prevent the application of research into practice. While this gap has been studied in multiple fields, it is unclear if findings apply to HCC, which may have unique challenges due to the nature of cybersecurity. Additionally, most gap research has focused on research outputs, largely ignoring potential benefits of research-practice engagement throughout the entire research life cycle. To address these gaps, we conducted a survey of 133 HCC researchers. We found that participants most often engage with practitioners during activities at the beginning and end of the research life cycle, even though they may see the importance of engagement throughout. This inconsistency may be attributed to various challenges, including practitioner and researcher constraints and motivations. We provide suggestions on how to facilitate meaningful researcher-practitioner interactions towards ensuring HCC research evidence is relevant, available, and actionable in practice.

1 Introduction

Human-centered cybersecurity (HCC) (also known as *usable security*) involves the social, organizational, and technological influences on people's understanding of and interactions with cybersecurity [43, 56]. Taking a human-centered approach to cybersecurity is critical given the significant role of human behavior in cyber incidents [3, 32, 74]. Yet, poor usability and over-reliance on technology to solve cybersecurity problems have led to frustration, anxiety, confusion, or complacency among both cybersecurity non-experts and experts [11, 45, 54].

The HCC research community endeavors to better understand and overcome these challenges, with an ultimate goal of facilitating human-centric design and implementation of cybersecurity technologies and processes that result in positive experiences and outcomes [43, 56]. HCC research can greatly benefit practice. For example, catalyzed by HCC password research [15, 65, 71], a revision of widely-adopted, practitioner-developed digital identity guidelines shifted burden (e.g., frequent password changes) away from end users, thus improving user authentication experiences [55]. Research on internet of things security and privacy labels [12, 28] informed the layered label approach of the new U.S. Cyber Trust Mark [30].

However, these examples are not the norm. Research and practitioner concerns may be out of sync, resulting in research with low likelihood of practitioner uptake [24]. Even when practitioners see the value of HCC, they may struggle to know how to implement HCC principles into their work [41], so fail to to effectively address the critical human component of cybersecurity [57]. To remedy these issues, it is imperative to encourage stronger connections between HCC research and practice [24, 41].

Unfortunately, research efforts in diverse disciplines have found that interests, incentives, values, and work routines of practitioners and researchers diverge in ways that make meaningful integration and collaboration a challenge [4, 6, 10, 44]. These disconnects, known as the *research-practice gap*, can adversely impact both communities [9, 10, 23, 44]. Practitioners may not benefit from research insights that could advance their work. Researchers may not benefit from practitioners' insights that could inform the pursuit of research meaningful and actionable to practitioners [17].

To date, there has been little investigation of the research-practice gap in the HCC field. Therefore, it is unclear if prior findings are applicable to HCC, given that cybersecurity is often cited as uniquely challenging because of its rapidly evolving technology and threats, adversarial setting, and so-

ciotechnical implications [22, 25, 31, 60]. While a prior survey explored practitioner perspectives on HCC research-practice integration [41], the researcher perspective is missing. Moreover, research-practice gap research has typically focused on activities at the culmination of research efforts (e.g., writing and distributing outputs) [18, 66], seldom addressing potential benefits of practitioner engagement from the "beginning to the end of the knowledge-creation process" [6]. Without an understanding of researcher-practitioner interactions in HCC, solutions to promote the integration of HCC research into practice cannot be developed and people will continue to struggle in their cybersecurity interactions. To remedy this, we conducted an online survey of 133 HCC researchers to answer the following research questions:

RQ1: What are HCC researchers' perceptions and experiences engaging with and considering practitioners and practitioner resources throughout the research life cycle?

RQ2: How do HCC researchers share research evidence with practitioners?

RQ3: What are barriers to practitioner engagement, if any?

RQ4: How do HCC researcher experiences differ, if at all, based on practitioner demographics?

We found that participants most often engage practitioners at the beginning and end of their research. Although they see the importance of engagement in most research activities, they do not always do so as they experience a high level of challenge. We identify a variety of challenges to these interactions, including a perceived lack of practitioner interest and researchers not knowing how to best engage.

Our study makes several contributions. We extend existing research-practice gap literature to provide domain-specific evidence valuable to the HCC community. Further, we provide the HCC researcher perspective, which can help identify disconnects in relation to HCC practitioner research [41]. We uniquely explore researcher-practitioner interactions across the entire research life cycle, providing novel insights into research activities that could benefit from increased interactions to ensure research is practice-appropriate and relevant from the start. We also identify interaction challenges and provide suggestions that can help researchers engage with practitioners and alleviate the burden currently placed on both communities. Lastly, we recommend future research that can extend our results and identify viable solutions for the HCC research-practice gap, ultimately working towards "important research" that "meets the needs of practice by addressing a real-world problem in a timely manner" [24].

2 Related Work

2.1 Research-Practice Challenges

Literature in diverse disciplines (e.g., social work [23], human-computer interaction [5, 17, 35], business [4, 6, 9], and con-

servation [44]) identify challenges that hinder research from making an impact on practice. Most focus on challenges at the end of the research life cycle: translation and sharing of research outputs. While researchers often carry the burden of knowledge translation, they do not always have credibility with practitioner audiences, the skills and experience to translate in formats and language understandable to practitioners, or time and resources [4, 17, 38, 44]. Further, academic researchers are often incentivized by obtaining a degree or tenure, which are dependent on producing novel contributions and publishing in academic forums. Therefore, they may not expend effort transferring their research into practitioner-focused formats [4, 17, 44]. Yet, some criticize practitioners for using low-quality or no research in their practice [4] or for misinterpreting research results [66]. In reality, practitioners may lack access or time to read research papers not in a format understandable to them [9, 17] and may not view research publications as timely given long publication timelines [4, 17]. Since practitioners are focused on maintaining daily operations or making a profit, they may be hesitant to change their processes to incorporate research recommendations when return on investment may be unclear [9, 47]. They may also not know how to apply research findings due to non-actionable or non-transferable recommendations [4, 5, 14, 17] or the theoretical nature of some research [8, 33].

Beyond research outputs, most challenges pertain to lack of cross-community communication and understanding. Practitioners rarely communicate their ideas about problems of most interest because there are few avenues for them to do so [4, 5, 17, 35]. This may result in the selection of research topics not compatible with practitioners' needs [4]. Additionally, researchers who lack practitioner experience themselves can have inaccurate or incomplete abstractions of practice that compromise the validity and applicability of their results [4, 35]. While practitioner resources, such as industry reports or technology news articles, can provide insights into practitioner contexts, researchers may be hesitant to use these [68]. Academic standards depend upon reliability, validity, and analyses as prerequisites for publishing in peer-reviewed journals. In contrast, practitioner publications may rely on case study examples with organizational viewpoints, have undisclosed methodologies and measures, focus on practical rather than theoretical implications, or place emphasis on emotion rather than facts [68].

2.2 Human-Centered Cybersecurity

While the research-practice gap exists in multiple fields – including the closely-related human-computer interaction (HCI) field – it is uncertain how manifestations of the gap in HCC may differ due to distinctive characteristics of cybersecurity. To start, cybersecurity exists in an ***adversarial setting*** [51, 60]. Adversaries are not just limited to malicious actors, but, of particular import to HCC, can also include end users viewed

as "enemies" or "the weakest link" [2, 36, 67, 75]. Cybersecurity is also characterized by its ***rapid pace of change*** with constantly evolving threats, technologies, and regulations [25, 26, 60]. Therefore, keeping up with the latest developments can be difficult for practitioners and researchers alike [22, 25, 51, 63]. Further, the ***intangible, uncertain nature*** of cybersecurity impacts, victims, and threats can hamper accurate assessments of risks, possibly leading to failure to act [22, 67]. Cybersecurity's uncertainty and dynamism result in ***contested debate about which solutions are most effective*** and how to show return on investment [22, 25, 63]. Cybersecurity researchers, in particular, are challenged to demonstrate definitive, reproducible results in the presence of myriad confounding variables [13]. Finally, and of particular HCC relevance, cybersecurity involves complex, ***sociotechnical relationships*** [22, 52, 60]. However, practitioners often take a techno-centric approach and may not be well-versed in human factors influences [21, 57].

To the best of our knowledge, only two studies addressed the research-practice gap in cybersecurity. One was focused on research topics [24], and the other, while looking at HCC, took a practitioner perspective [41], leaving the researcher perspective unknown. Further, existing research-practice gap literature provides a limited view of researcher-practitioner engagements, with none exploring interactions across the entire research life cycle. These shortfalls leave the HCC research community unsure about when is most advantageous to engage practitioners and how to ensure their research evidence is relevant, available, and actionable for practitioners to leverage. Our study begins to address these gaps.

3 Methodology

To explore researcher engagement with practitioners, we conducted an anonymous, online survey of 133 HCC researchers in July 2023. Our Research Protections Office approved the study. On the first survey screen, we provided information about participant rights and data protection. Participants did not receive monetary compensation. Responses were anonymous and assigned identifiers (e.g., R10).

3.1 Survey Development

We selected a predominantly quantitative survey study design since existing qualitative literature (e.g., [9, 14, 17, 24, 66]) served as a foundation for developing survey questions and responses and we wanted to gauge the prevalence of those findings within HCC. Further, the survey format afforded identification of areas of interest that could be targeted in future HCC-specific studies. Two subject matter experts reviewed an initial draft to check for clarity and completeness. Each reviewer had over 20 years of experience conducting usability and HCC research, and one had prior practitioner experience. We adjusted the survey instrument based on their feedback.

The final survey (Appendix A) consisted of select-one-option, select-all-that-apply, Likert scale, and open-ended questions.

3.1.1 Topics

The first survey section collected professional demographic information. Participants then indicated the frequency, perceived importance, level of challenge, barriers, and methods of consulting practitioners (obtaining input directly from practitioners, e.g., via email or in-person) or practitioner resources during various research activities. We aligned the activities with research life cycle phases for which practitioners or practitioner resources could potentially be consulted [76]: Research Conceptualization; Study Design; Data Collection; Data Analysis; and Dissemination.

3.1.2 Terminology

To ensure participants had a common understanding of HCC and practitioners, we described each term at the beginning of the survey (Appendix A). Since there is no standard definition for ***human-centered cybersecurity*** or the related term usable security [70], we created a description based on explanations from other HCC research groups [39, 43, 56, 72]. Our description of ***practitioners*** was largely informed by a prior narrative on security information workers [77]. Examples of practitioners include: cybersecurity practitioners, such as analysts, architects, and consultants; IT practitioners, such as administrators, help desk, system and network architects; developers; organizational leadership; policy makers; and cybersecurity educators and trainers. When asking about practitioner engagement, we also included consultation of ***practitioner resources*** (e.g., industry and government reports, technical standards and guidelines, and policies) since these can serve as a proxy for practitioner perspectives.

3.2 Survey Data Collection and Participants

Eligible participants had to be adults (18+ years of age) and have experience conducting HCC research. To recruit participants, we sent email invitations to a compiled list of authors of HCC papers published the prior three years at applicable conferences (e.g., Symposium on Usable Privacy and Security, USENIX Security Symposium). The full list of conferences is in Appendix B. We also emailed professional contacts and advertised via social media posts and a cybersecurity mailing list. The survey, implemented on the Qualtrics platform, was open for three weeks. During a data quality check, we excluded partial responses and responses where participants indicated they were not researchers. We also looked for abnormally low completion times and nonsensical open-ended responses (not finding either), finalizing on 133 survey responses.

Table 1 shows participant demographics. The largest percentage were tenure-track/tenured faculty, followed by graduate students. The majority (75%) had 10 or fewer years of

Table 1: Participant demographics (N = 133)

Demographic	Response Option	n	%
Research position	TT* faculty	53	39.85%
	Non-TT* faculty	14	10.53%
	Graduate student	41	30.83%
	Other researcher	25	18.80%
Practitioner experience	Yes, currently	32	24.06%
	Yes, in the past	59	44.36%
	No	42	31.58%
Years of experience	Less than 1	7	5.26%
	1 to 5	51	38.35%
	6 to 10	42	31.58%
	11 to 15	19	14.29%
	16 to 20	8	6.02%
	More than 20	6	4.51%
Organization type	Academic	106	79.70%
	Private industry	14	10.53%
	Non-profit	4	3.01%
	Government	8	6.02%
	Other	1	0.75%
Region	Africa	4	3.01%
	Asia	5	3.76%
	Europe	56	42.11%
	North America	66	49.62%
	Oceania	2	1.50%

* TT = tenure-track

experience conducting HCC research. Most worked in an academic institution. Ninety-two percent worked in North America or Europe. Sixty-eight percent indicated that they had been or currently were practitioners. Of those participants, 45% had security practitioner experience, 43% indicated developer experience, 29% had been educators/trainers, 25% had been IT practitioners, 23% had management experience, 9% were policy makers, and 7% indicated "Other."

Participants reported the user populations that have been the focus of their HCC research, and then the populations that could make use of or put into practice the implications and recommendations from their research (Fig. 1). Participants most often studied general public end users (71%), followed by organizational end users (50%) and security practitioners (47%). Among those who selected Other, vulnerable and at-risk populations (e.g., individuals with disabilities, children, the elderly) were the most mentioned, so are specifically included in the figure. Populations who can make use of participants' research were much more evenly distributed, with over half selecting all but three populations. The most-selected were security practitioners (74%) and policy makers (71%). Only nine did not select a practitioner group.

3.3 Survey Data Analysis

We calculated descriptive statistics and inferential statistics at a significance level of $\alpha = 0.05$ to explore differences across the data. We also conducted qualitative data analysis for the one open-ended survey question.
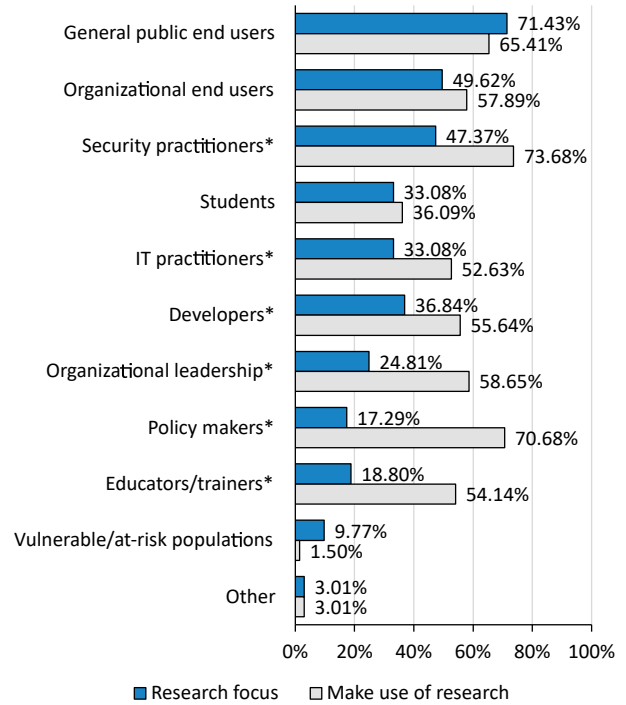
Figure 1: Population focus of research and populations who could use research (n = 133). * practitioner population.

3.3.1 Statistical Analysis

We compared independent groups for three participant demographic variables with potential to influence responses, combining several demographic groups for greater statistical power. Based on literature suggesting that researchers with non-academic experience engage in more external interactions [40] and that "pracademics" (those with both academic and practitioner experience) are useful in bridging the theory-practice gap [59], we posited that prior or current **practitioner experience** might influence participants' experiences and views about engaging practitioners during research activities. Practitioner experience consisted of two groups: those with prior or current experience as a practitioner (n = 91) and those without practitioner experience (n = 42). We were also interested in the impact of **organization type** since institutional incentives were found to be a factor in prior research-practice gap literature [44]. Groups included academic (n = 106), private industry (n = 14), and "other," primarily consisting of participants from non-profits and government (n = 13). Finally, we tested the impact of prior experience conducting **practitioner-focused research** since connections made during this research may afford researchers the ability to enlist practitioner support for future efforts [64]. We considered participants with practitioner-focused research (n = 96) to be those who indicated at least one practitioner population in Fig. 1. All others were in the "no practitioner-focused research"

group (n = 37).

To compare ordinal (Likert scale) responses for variables with two independent groups (e.g., practitioner experience), we used Mann Whitney U tests, reporting significant results with the z-statistic. For the three groups of organization type, we performed an initial Kruskal Wallis H test with a post-hoc Dunn's test adjusted for multiple comparisons using the Holm-Šidák correction [1], reported with z. We also report the effect size, Cohen's d, with the following thresholds: small 0.20; medium 0.50; and large 0.80 [16]. A medium or large effect size may indicate that a finding has practical significance [69].

For categorical question responses, we used Chi-square tests of association – reported with χ^2 (one degree of freedom) – or Fisher's exact tests in instances of five or less occurrences [46]. We report the effect size, Cramer's V. For one degree of freedom, small, medium, and large effect size thresholds are 0.10, 0.30, and 0.50, respectively [46].

Note that an *a priori* power analysis [29] for a Mann-Whitney U test with similarly-sized groups (medium effect size, $\alpha = 0.05$, power = 0.8) yielded a minimum sample size of 134, while we had 133. Because of challenges recruiting this specialized population and unevenness of group sizes due to convenience sampling, we acknowledge that statistical power may be lacking, thus creating a risk of not finding a difference that is actually there [37].

3.3.2 Open-ended Question Analysis

We employed qualitative coding techniques to analyze responses from an open-ended question at the end of the survey asking participants if they had additional thoughts. Two research team members first individually read through the responses and developed an initial set of codes. They then met to discuss and decide on a codebook. Since the data set was small (30 responses averaging 45 words per response), there were only five codes (see 4.6.3). The two researchers then independently coded all responses using the codebook and met again to discuss and resolve the few coding differences.

4 Results

We report summary statistics and significant inferential statistical results. The absence of significant result reporting for a question signifies there were no differences for any variable of interest. We organize this section by research phase. Figures 2, 3, and 4, referred to throughout, show the frequency, importance, and challenge ratings, respectively, for research activities. Frequency responses were on a 5-point scale ranging from never to always. Importance was rated on a 5-point scale from not important to extremely important. Level of challenge was on a 5-point scale from extremely challenging to not challenging with a "no experience to judge" option.
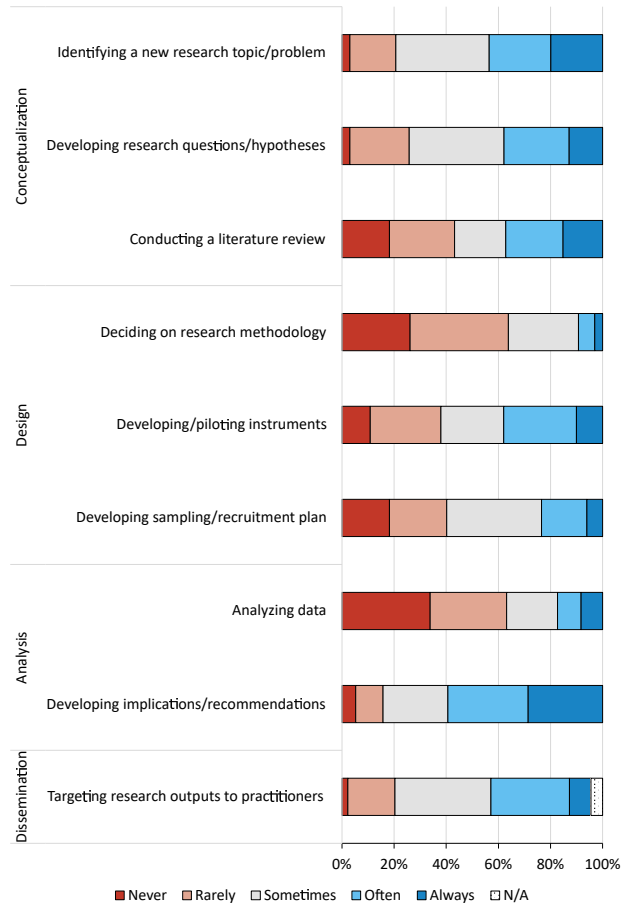
Figure 2: Frequency with which participants consult practitioners/practitioner resources during research activities. N/A only applies to "Targeting research outputs to practitioners," indicating a participant does not produce research outputs.

4.1 Research Conceptualization Phase

Participants answered questions about three activities within the research conceptualization phase.

Identifying a new research topic or problem. Less than half (44%) said they consult practitioners or practitioner resources often or always when identifying a new research topic, with 21% selecting never or rarely (Fig. 2). Seventy percent said consulting practitioners was moderately or extremely important for this activity (Fig. 3). Forty-two percent said that practitioner consultation had been moderately or extremely challenging (Fig. 4).

Developing research questions or hypotheses. Just 38% of participants said they often or always consult practitioners when developing research questions or hypotheses, and 26% said they rarely or never do (Fig. 2). Over half (56%) said that it was moderately or extremely important to do so (Fig. 3). About three-quarters (76%) indicated that it was at least somewhat challenging to consult practitioners during this
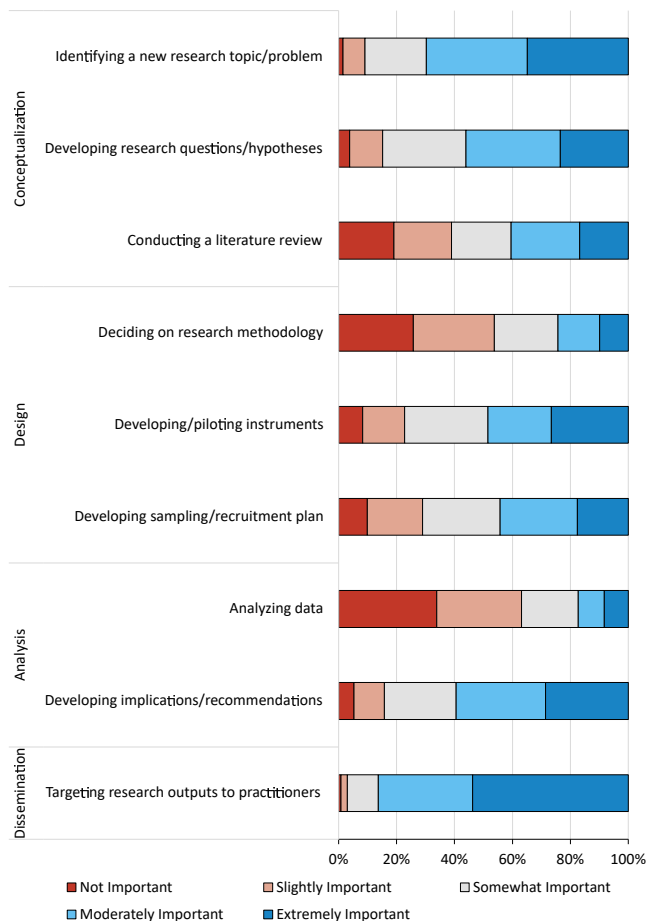
Figure 3: Perceived importance of consulting practitioners/practitioner resources during research activities

Figure 4: Level of challenge consulting practitioners/practitioner resources during research activities. Participants who did not recruit practitioners are counted as "no experience to judge."

activity (Fig. 4).

Conducting a literature review. Thirty-seven percent consult practitioners/practitioner resources often or always when conducting a literature review (Fig. 2). They generally viewed consultation during this activity as less important (40% not/slightly important) (Fig. 3). Only 26% found this to be moderately/extremely challenging, with 19% not having the experience to judge (i.e., they had never attempted it) (Fig. 4).

Demographic differences. When *identifying a research topic*, industry participants consulted practitioners significantly more often as compared to those working in academia and other organizations ($z = 3.05$, $d = 1.31$). Additionally, participants who had conducted practitioner-focused research more frequently consulted practitioners ($z = 2.67$, $d = 0.48$) and rated consultation higher in importance ($z = 2.64$, $d = 0.50$) compared to those who had not. When *developing research questions*, industry participants consulted practitioners more often than those in academia ($z = 2.63$, $d = 0.79$)

4.2 Study Design Phase

We asked participants questions related to three activities in the study design phase.

Deciding which research methodology is most appropriate. Only 9% of participants often/always consult practitioners when deciding on research methodology, with 64% selecting never or rarely (Fig. 2). Few thought practitioner consultation was important (24% moderately/extremely important, 54% not/slightly important) (Fig. 3). Twenty-eight percent said it was moderately/extremely challenging, and 30% had no experience to judge (Fig. 4).

Developing and piloting research instruments or experiments. A minority (38%) often/always consult practitioners when developing and piloting their research instruments or ex-

periments, with the same percentage never/rarely consulting (Fig. 2). A higher percentage (48%) said consulting during this activity was moderately/extremely important (Fig. 3). Participants expressed a fair amount of challenge for this activity, with 42% saying it is moderately/extremely challenging to consult practitioners (Fig. 4) and 16% indicating they had no experience to judge.

Developing a sampling or recruitment plan. Few participants (24%) indicated that they often/always consult practitioners when developing a sampling or recruitment plan, with 40% selecting never or rarely (Fig. 2). More thought consulting during this activity was important (44% moderately/extremely important) (Fig. 3), although challenging (42% moderately/extremely challenging) (Fig. 4). Thirty percent said they had no experience to judge the challenge.

Demographic differences. For *developing/pilot research instruments*, participants who had conducted practitioner research more frequently consulted practitioners (z = 3.35, d = 0.68) and had higher importance ratings (z = 3.21, d = 0.62) compared to those who had not. For this same activity, importance ratings from participants in other organizations were significantly higher than ratings from those in academia (z = 2.59, d = 0.77) and industry (z = 2.25, d = 0.98). When *developing a sampling/recruitment plan*, those who had conducted practitioner-focused research consulted practitioners significantly less frequently (z = -2.22, d = 0.44).

4.3 Data Collection Phase

We had a different vein of questioning for the one activity, recruiting practitioners, in the Data Collection phase since not all researchers enlist practitioners as research subjects. Therefore, questions related to frequency and importance were not applicable for this activity.

Among the 68% who had recruited practitioners, interviews (63%) and surveys (61%) were the most common study types, with 32% recruiting practitioners for experiments, 32% for focus groups/workshops, and 14% for another purpose. These participants were asked two additional questions.

Recruitment methods. Professional contacts and snowballing were the most popular methods of recruiting practitioners (Fig. 5). Among those who selected "Other," conferences and events (5 participants) and freelance platforms such as UpWork (4) were most mentioned. Other recruitment mechanisms included GitHub, Discord/Slack, contacting practitioners mentioned in online articles and websites, and soliciting participants from prior studies.

Recruitment challenge. Most (72%) indicated that recruiting practitioners was moderately/extremely challenging, with only 3% saying it was not challenging (Fig. 4).

Demographic differences. For *recruitment methods*, participants in academia were more likely to select snowballing as compared to those in industry (χ^2 = 6.21, V = 0.25).
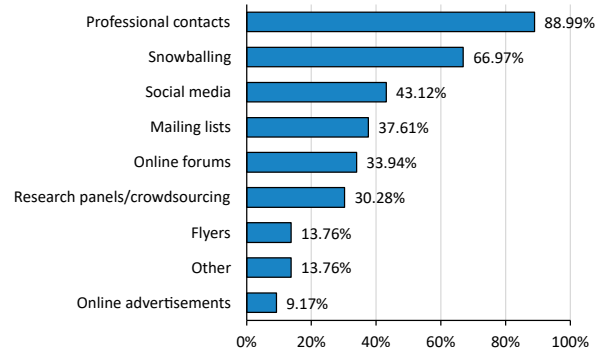
Figure 5: Practitioner recruitment methods (n = 109)

4.4 Data Analysis Phase

Participants answered questions related to two activities within the data analysis phase of research.

Analyzing data. Participants infrequently consult practitioners when analyzing data, with 64% selecting rarely or never (Fig. 2). This was also reflected in importance ratings, with 63% rating consultation as not/slightly important (Fig. 3). Only 22% were moderately/extremely challenged consulting practitioners during this activity (Fig. 4), with an appreciable number (41%) indicating they had no experience to judge.

Developing implications, recommendations, and solutions. Participants frequently consult practitioners when developing implications, recommendations, and solutions (59% often/always) (Fig. 2) and believe doing so to be important (59% moderately/extremely important) (Fig. 3). Thirty-four percent were moderately/extremely challenged, with 22% indicating they had no experience to judge (Fig. 4).

Demographic differences. Academic participants rated the challenge during the *analyzing data* activity significantly higher than those in other organizations (z = 2.42, d = 0.83).

4.5 Dissemination Phase

We asked participants several questions pertaining to the research output dissemination phase.

Producing or contributing to research outputs targeted at practitioners. Participants indicated the frequency with which their research outputs (e.g., papers, tools) are targeted at practitioners on a 5-point scale (never - always) with a "I do not produce or have not yet produced research outputs" option (Fig. 2). Just 38% said they often/always produce these outputs. However, 75% said they at least sometimes do. Most (86%) thought it was moderately/extremely important to produce these outputs (Fig. 3). Yet, over half (57%) indicated that this was moderately/extremely challenging to do (Fig. 4).

Research output dissemination channels. Participants who produced practitioner-targeted research outputs at least rarely (n = 124) were asked how they disseminate those (Fig. 6).
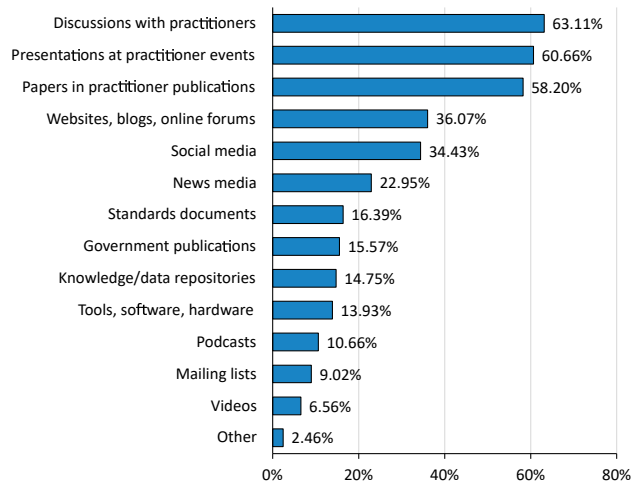
Figure 6: Channels through which research outputs are disseminated to practitioners (n = 124)

Over half selected discussions with practitioners, presentations at practitioner conferences, meetings, and events, and papers/articles in practitioner-focused publications. Two participants selecting "Other" indicated academic forums: *"publishing in academic places and hoping they'll see it" (R83)*.

Practitioner impact and interest. Participants indicated how often they think their research directly impacts practice, yielding the following responses: 3% never, 9% rarely, 38% sometimes, 20% often, 14% always, and 16% do not know. Researchers also selected the extent to which they believe practitioners would be interested in having research outputs shared with them: 2% reported not interested, 16% slightly interested, 29% somewhat, 39% moderately, and 14% extremely.

Demographic differences. Academic participants were less likely to select the government publication channel compared to those from other organizations (Fisher's exact, p = 0.010, V = 0.29). Those conducting practitioner-focused research more often selected the following: presentations at practitioner forums (χ^2 = 9.93, V = 0.29); tools, software, and hardware (Fisher's exact, p = 0.003, V = 0.25); and knowledge/data repositories (Fisher's exact, p = 0.02, V = 0.21).

4.6 Barriers

Participants selected barriers encountered when engaging with practitioners. Because the challenges encountered during the dissemination phase may not apply to other research phases, we asked separate questions about barriers encountered before and during dissemination. Further, in an open-ended question, participants shared additional thoughts about practitioner interactions, with all comments related to barriers.

Figure 7: Barriers to consulting practitioners/practitioner resources during pre-dissemination research phases.

4.6.1 Pre-dissemination Barriers

Practitioners not having time was the most selected barrier at 67% (Fig. 7). No other barrier was chosen by a majority. Over 40% indicated that organizations do not allow practitioners to participate, practitioners do not see the value in participating, and they are not sure how to reach practitioners. Only 19% said they have little or no incentive to consult practitioners/practitioner resources, and just 8% said they do not have time. Among the write-in responses for "Other" barriers were: practitioners being wary or thinking they are being audited (3 participants); uncertainty about whether it is appropriate to cite practitioner resources; and inadequate financial incentives for practitioners to participate. Participants who had conducted practitioner-focused research less often selected little or no incentive (χ^2 = 8.78, V = 0.26) and more often selected practitioners not having time to participate (χ^2 = 4.18, V = 0.18) as compared to participants who had not.

4.6.2 Dissemination Barriers

Dissemination barriers varied, with no individual barrier selected by a majority (Fig. 8). The most selected was lack of interest or uptake from practitioners (41%). Over 30% said that there was little funding or resources, they were not sure where to disseminate results, there was little or no incentive, and that they were not sure how to translate research into content valuable to practitioners. Write-in responses for "Other" included: women's work not taken seriously in male-dominated fields; practitioners wanting validated, replicated, and quantifiable results; and a language barrier. Those who
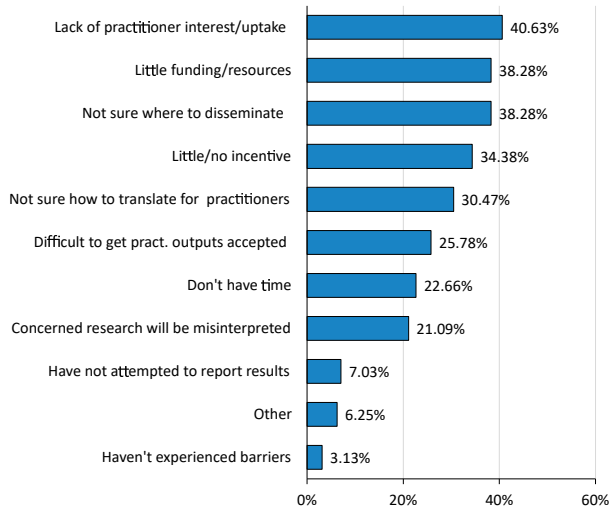
Figure 8: Barriers to producing or contributing to research outputs for practitioners.

had conducted practitioner-focused research were less likely to say there was little/no incentive (χ^2 = 4.30, V= 0.18).

4.6.3 Qualitative Comments

We identified five main barriers in 30 open-ended responses.
Difficulty making connections. Twelve participants offered comments about not knowing how to reach practitioners. For example, an academic shared their frustration: *"I often reach out to practitioners to discuss study designs or disseminate results. Most of the time, I never hear back" (R47)*. Several expressed uncertainty about where to find practitioners who might benefit from their research: *"It's hard from the outside to know which people, at which organizations, might be interested in the specific area that you work on" (R59)*. Recruiting practitioners, as found in quantitative results, is a particular challenge *"even after offering compensation" (R103)*. A lack of researcher or institutional name recognition may also hinder getting practitioners' attention: *"I believe they respond to requests from popular researchers but have no incentive in responding to a wider range of researchers" (R120)*.

Several proposed ways to facilitate contact: *"It would be great to have a forum for collaborating – identifying practitioners with interests that overlap mine" (R112)*. R120 similarly suggested, *"having an organization or forum that enable academics to 'pitch' their projects to practitioners in the hopes of getting them to participate will revolutionize human centered security research."* A faculty member in Europe called upon research funding institutes to *"act as a facilitator between academia and industry" (R76)*.

Divergent interests. Lack of practitioner interest in research may be due, in part, to conflicting interests and priorities of the two communities, mentioned by seven. Several expressed

uncertainty about whether HCC research efforts are valued by practitioners. R81, a graduate student, said, *"In my opinion practitioners are focused on business/profit and not on effectivity of interventions, thus are not interested in HCI/security research."* Others commented that HCC research topics may not align with areas of practitioner interest. For example, R100 remarked, *"the issues the academic community values, e.g. privacy, are not necessarily valued by practitioners."*

Practitioner hesitance to share data. Four participants mentioned organizations being hesitant to share sensitive cybersecurity data. One stated, *"The most difficult challenge I face in working with practitioners is getting approval (from their organizations) to share security related data with external parties" (R27)*. An academic similarly commented, *"Practitioners are concerned of exposing their security loopholes" (R39)*. Another shared an example in which they were unable to address a discovered security issue because a business was reluctant to share data: *"We identified a vulnerability across a diverse pool of practitioner groups, but they were DISincentivized from communicating openly with us due to fears of opening themselves up to liability" (R24)*.

Lack of researcher incentives and time. Three commented on lack of incentive and time. While a graduate student felt researchers should *"talk more to people who actually do the things we just theoretically discuss" (R03)*, they lamented that in *"publish or perish academia, in which I hopefully gather multiple top tier conferences to graduate my PhD in a few years, I just don't have time to even think about doing additional projects or publications for practitioners."* A tenure-track faculty noted lack of institutional support: *"these [practitioner] publications do not contribute towards my academic promotion, so there is little incentive" (R116)*.

Presentation challenges. Nine participants cited difficulties presenting results in a way that is meaningful to practitioners. The sometimes abstract or non-generalizable nature of research findings poses challenges for researchers when trying to provide takeaways and recommendations. For example, R80, a North American academic researcher stated, *"As study results aren't always 'clean,' communicating the nuance of research findings to practitioners while providing useful, actionable insights can be challenging."* A European researcher commented that practitioners *"demand simple answers for very difficult questions" (R53)*. Researchers may further struggle to develop interpretations of research evidence actionable by practitioners since they do not understand the practitioner context. A participant with practitioner experience stated, *"I often see research in this space aimed at practitioners that don't understand their perspectives well, and present fairly naive/superficial results" (R59)*. Additionally, producing outputs in a style appropriate to the constraints and needs of practitioners can be non-trivial, as expressed by R92: *"I find that the challenge of writing for a different audience is difficult for fellow researchers without prior experience in industry."*

5 Discussion

Our study provides novel insights into the research-practice gap within the HCC field, specifically how researchers currently engage practitioners throughout the research life cycle and the challenges they encounter. In this section, we discuss limitations, revisit our research questions, provide practical recommendations, and suggest future research opportunities.

5.1 Limitations

We acknowledge several study limitations. There may be self-selection bias as those choosing to participate might have an interest in the survey topic and may not represent other researchers' views. Additionally, while we recruited researchers publishing in international conferences, most venues largely featured papers from North America and Europe, as reflected in participant demographics. Therefore, it is unclear if our results transfer to other regions since institutional incentives may differ [49]. Finally, our largely quantitative survey did not explore reasons behind responses. Using our study as a foundation, we recommend additional, qualitative research to further explore areas of interest identified in the survey as well as lessons learned in research-practice interactions and potential solutions.

5.2 Understanding Interactions

In this section, we discuss our findings in relation to our research questions as well as HCC-specific insights.

RQ1: Researchers see the value of engaging with practitioners but are often challenged in doing so. Our participants recognize that practitioners are key to their research making an impact (Fig. 1). The majority viewed connections with practitioners as at least somewhat important during most research activities and particularly so in activities at the beginning (e.g., identifying a new research topic) and end (e.g., targeting research outputs to practitioners) of the research life cycle. However, as our results illustrate, researchers are often highly challenged to connect with practitioners, so interactions may not actually happen.

RQ2: Dissemination channels do not always match practitioner preferences. Participants disseminate practitioner outputs via a variety of channels, most often through conversations with practitioners. In comparing these channels to those practitioners prefer [41], we see that while both communities favor presentations and articles in practitioner forums, there are substantial gaps for other channels. Compared to practitioner preferences, our participants more often share their results via researcher-practitioner discussions, social media, and news media. Researchers less often share their outputs via websites, standards documents, government publications, knowledge/data repositories, tools, podcasts, mailing lists, and

videos. The differences indicate a current disconnect but also a roadmap for where researchers can invest more effort.

RQ3: Barriers differ across the life cycle. We uniquely identify researchers' challenges pre-dissemination, finding that these are often dependent on practitioner context (e.g., practitioner time, perception of research value, and organizational gate keeping) rather than issues on the researcher end. Conversely, barriers encountered during dissemination more often reflect issues in the research context that are similarly cited in existing literature (e.g., lack of resources, time, incentives, and translation knowledge [9, 38, 44, 47]). We extend this prior research by quantifying the frequency with which these barriers are encountered in HCC, finding none were selected by a majority. Further, lack of incentive and time, which are frequently cited as major challenges during knowledge translation and sharing [4, 17], were selected by a minority. Although our different results may be influenced by self-selection bias, they may also signify a shift towards HCC researchers becoming motivated to influence practice.

RQ4: The differences across demographic groups are likely due to access and opportunity. While we anticipated that participants with prior *practitioner experience* would interact more with practitioners and be more likely to see the importance in doing so, there were no statistically significant results to support this. Given this unexpected result, we suspect there may be other factors at play, for example, recency of practitioner experience or relevance to the research. Additionally, a potential lack of statistical power (described in 3.3.1) may also explain the lack of significant results. Divergences across *organization types* might be due to differing levels of access to practitioners. For example, industry and government participants may have ties to practitioners within their organizations, so would be more likely to consult them when identifying a research topic and have less need to use snowball recruitment. Differences among *practitioner-focused research* groups are likely due to the nature of the research; practitioner-focused research naturally necessitates more interactions. This was evident in differences consulting practitioners at the beginning of the research life cycle, dissemination of research outputs in ways preferred by practitioners [41], and incentive to engage practitioners. While these findings are not surprising, we see a missed opportunity for researchers not conducting practitioner-focused research since practitioners are often the designers of technologies that cause issues for end user populations and the ones to ultimately implement researchers' recommendations.

Domain-specific insights: Our results may reflect the distinctive characteristics of cybersecurity. In a contested field already challenged to prove return on investment [25, 63], cybersecurity practitioners may be reticent to embrace research not proven in an operational context or without concrete recommendations [25, 41, 53, 61]. Further, because many practitioners are technology-oriented, they may not value sociotech-

nical considerations [52,60]. Challenges of a constantly evolving, uncertain field result in practitioners being overworked and burnt out [27,58], and, therefore, less willing to spend time reading or participating in research [7,25,41,64]. Moreover, within an adversarial setting – not present in the related field of HCI – hesitance to disclose sensitive cybersecurity information and distrust of researchers [7,25,62] may lead to organizations not allowing their employees to participate.

We observe evidence of similar research-practice challenges within the adjacent domain of human-centered artificial intelligence (AI), which shares with cybersecurity characteristics of fast-paced change, focus on technology solutions, and sociotechnical entanglements [19,42,48,73]. Research efforts on AI ethics (e.g., fairness and privacy) from the perspective of AI practitioners found some research-practice challenges similar to those in HCC, for example, reluctance to share sensitive data, lack of motivation to advocate for human element considerations (e.g., privacy) when return on investment is uncertain, lack of awareness of the severity of human-centric threats, and a disconnect between practitioner needs and solutions provided by researchers [42,48,73].

To overcome the disconnect, recommendations for AI researchers center on building trust through practitioner collaborations (e.g., conducting studies in real-world contexts), not just to understand current problems in practitioner processes but also to work towards fixing those [73] by providing tools, frameworks, checklists, and "integrative approaches that address awareness, motivation, and ability together" [48]. These recommendations can also apply to HCC in addressing the challenges found in prior HCC practitioner research [41] and our study. In turn, our research can contribute a researcher perspective, identifying advantageous points of researcher-practitioner interactions throughout the life cycle and researcher-specific challenges that may be applicable to human-centered AI. However, we see the need for more research to delve deeper into not just the commonalities, but also how the differences across and at the intersection of the two domains (e.g., the relative newness of AI implementations in practice vs. those in a more mature cybersecurity field) may impact research-practice challenges and solutions.

5.3 Practical Implications

We offer suggestions towards bridging the research-practice gap. While directly linked to our results, we recognize the need for further work to determine feasibility and acceptance of these for the research and practice communities.

5.3.1 For Researchers

Consider where additional practitioner engagement might be beneficial. Consulting practitioners/practitioner resources early and often during the research life cycle can help ensure that research is relevant to practitioner needs and context.

While many participants saw the value of consulting practitioners at the beginning and end of the life cycle (Fig. 3), we suggest that, in some situations, there may be benefits to engaging practitioners during other activities. When conducting a literature review (4.1), despite possible researcher hesitance [68], the use of authoritative, credible practitioner resources (e.g., government publications, industry data breach reports, market analysis) could be helpful in identifying current cybersecurity issues and trends. In the study design phase (4.2), it might be valuable to ask practitioners representative of or familiar with users in the target study population for feedback on method appropriateness (e.g., whether an interview or focus group might be more acceptable to participants), survey instruments and design (e.g., feedback on technical accuracy/language, coverage of the topic, or completion time), or ways to recruit participants. While practitioners may not be well-versed in analysis (4.4), consulting them or practitioner resources during this activity may be helpful to better understand the context and meaning of data, for example, technical jargon and references in qualitative comments. These understandings can lead to the development of recommendations more relevant and actionable to practitioners.

Meet practitioners where they are. Some participants struggled with knowing where to disseminate outputs to practitioners (4.6.2, 4.6.3). We suggest they shift their efforts towards the information channels most preferred by practitioners (5.2). To accommodate researchers' time and resource constraints (4.6.2, 4.6.3), several channels require lower levels of effort, for example, being a guest on an established podcast or writing a short blog. These may also allow researchers to summarize key takeaways in their own words to avoid misinterpretation of results, an issue identified by researchers when interacting with news media [66]. Building a network of practitioners on social media (vs. posting to a following of mostly researchers) or joining mailing lists and forums that are frequented by practitioners can facilitate advertisement of outputs. Additionally, it may be beneficial to build relationships with science communicators, who are skilled in translating research information to practitioner terms and can share curated research evidence via channels practitioners prefer [18,66].

Determine the best time to report. In addition to knowing where to disseminate results, knowing *when* may be just as important. Practitioners, who may not always trust or see value in research results [4,41], might be more willing to act on conclusions originating from multiple studies, rather than from a single study. Therefore, we see a need for research synthesis reports, a model common within the medical profession [44].

Investigate ways to determine impact. Most participants were able to make a determination on how much impact their research has in practice (4.5), prompting us to wonder *how* researchers know this. Academic impact factors (e.g., H index) are not useful for measuring practitioner engagement. It may be difficult to determine if research is accessed, seen as

relevant, and implemented by practitioners. Therefore, we see value in future research that seeks to identify ways in which researchers gauge impact on practice and develop guidance on additional measures. These indicators of impact could provide encouragement to researchers, who, like our participants, may sometimes be disincentivized by institutional emphasis on research publications or are demotivated by a perception that practitioners are not interested in their work (4.6.2, 4.6.3).

5.3.2 For Intermediaries

While the above-mentioned strategies might help connect the two communities, they place the majority of the burden on researchers, who, as evidenced by our results, may lack the time, motivation, and skills to engage with practitioners or translate research into practitioner-friendly formats (4.6.1, 4.6.2, 4.6.3). Practitioners likewise struggle with similar issues (e.g., time, motivation) that keep them being more engaged with HCC research [41]. Therefore, it is important to explore solutions that alleviate undue burden on either community by enlisting the support of intermediary institutions and individuals.

Create space for research within practitioner forums. Our participants expressed challenges in knowing where to disseminate their findings and getting their outputs accepted to practitioner forums (4.6.2). To address these issues, intermediaries can feature HCC research in their events and publications. For example, conference sponsors could make a concerted effort to feature more research talks and offer grants to encourage researchers to attend. Cybersecurity organizations could invite researchers to present their work via channels that reach a broader practitioner audience, such as webinars or podcasts. Practitioner magazines and newsletters could include content featuring HCC research.

Instruct researchers on how to communicate to practitioners. To be able to effectively present research findings in the above-mentioned forums, researchers need to know how to create tailored, translational resources, ensuring outputs are actionable and prescriptive [10, 17, 20, 38, 61]. To develop these skills, which some participants indicated they lack (4.6.2, 4.6.3), educators of researchers can provide instruction on how to translate research findings to lay audiences for practical impact. Further, funding institutions can help researchers develop a business case and pitch to practitioners [50].

Establish evidence bridges. To reduce researchers' challenges (Fig. 4, 4.6), one proposed solution is the creation of *evidence bridges*, "professional individuals or organizations with a mandate to act as intermediaries between science and practice" [44]. These independent intermediaries synthesize and make accessible primary research in a format consumable by practitioners while providing a channel for practitioners to communicate their needs to researchers. These bridges are common in the medical field, for example, the American Cancer Society and Royal College of Physicians and Surgeons. To be successful, evidence bridges should have strong connec-

tions with and be trusted by both communities. A future investigation could help determine which current organizations, if any, may be best positioned to serve as evidence bridges. For example, we see a potential role for public research funding organizations (e.g., U.S. National Science Foundation, European Research Council) to assist their grantees in making impact in practice. Some universities have technology transfer organizations (e.g., [34]), which could be expanded to include the transition of research knowledge and recommendations. Additionally, since practitioners want HCC information from sources they trust or consider to be authoritative (e.g., standards documents and publications) [41], there may be a role for standards and government organizations to integrate HCC research insights in their technical publications.

Provide venues for researchers and practitioners to have meaningful interactions. While evidence bridges can be beneficial, not all communication should done via an intermediary. Compared to practitioner HCC interest ratings in a prior study [41], our participants underestimated practitioners' interest in receiving HCC information (4.5). They also expressed difficulty making connections with practitioners (4.3, 4.6.1, 4.6.2, 4.6.3). These findings indicate a communication gap and a need for improvement in relations. To facilitate dialogue, it is important to have venues where practitioners and researchers can meet to engage in meaningful discussion and begin to create connections for future interactions. These venues – perhaps organized by intermediaries – could move beyond one-way communication (presentation) formats towards a more interactive setting that allows for the mutual exchange of ideas and feedback. This exchange could shape future research, help researchers understand practitioner contexts, and provide practitioners with awareness of HCC.

6 Conclusion

Given the role of human error in cyber incidents, there is a critical need for practitioners to better address the human element. However, HCC research insights that could help advance their efforts may not be utilized, in part due to lack of access, relevance, and actionability, illustrating a research-practice gap [41]. Thus, informing researchers about how their efforts can better serve practitioner needs is key. Towards understanding and reducing the research-practice gap in HCC, we surveyed 133 HCC researchers on how they engage with practitioners. We extend existing knowledge by uniquely exploring the HCC researcher perspective and researcher-practitioner interactions across the entire research life cycle. We provide suggestions on facilitating integration of HCC research into practice by incorporating practitioner needs and context throughout the research process and enlisting intermediaries to connect the two communities.

Acknowledgements

We would like to thank the anonymous reviewers, especially our shepherd, for their efforts and comments that helped improve the paper. Special recognition also goes to our colleagues Mary Theofanos and Steven Furnell who provided valuable input to the study. Finally, we are incredibly grateful to the human-centered cybersecurity researchers who took time from their busy schedules to take the survey.

Disclaimer

Certain commercial companies or products are identified in this paper to foster understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the companies or products identified are necessarily the best available for the purpose.

References

[1] Hervé Abdi. Bonferroni and šidák corrections for multiple comparisons. *Encyclopedia of Measurement and Statistics*, 3(01):2007, 2007.

[2] Anne Adams and Martina Angela Sasse. Users are not the enemy. *Communications of the ACM*, 42(12):40–46, 1999.

[3] Arctic Wolf Networks. State of cybersecurity: 2022 trends. https://arcticwolf.com/resource/aw/the-state-of-cybersecurity-2022-trends/, 2022.

[4] Catherine Bailey. Employee engagement: Do practitioners care what academics have to say–and should they? *Human Resource Management Review*, 32(1):100589, 2022.

[5] Jordan Beck and Hamid R Ekbia. The theory-practice gap as generative metaphor. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2018.

[6] Inge Bleijenbergh, Jorrit van Mierlo, and Tanya Bondarouk. Closing the gap between scholarly knowledge and practice: Guidelines for HRM action research. *Human Resource Management Review*, 31(2):100764, 2021.

[7] David Botta, Rodrigo Werlinger, André Gagné, Konstantin Beznosov, Lee Iverson, Sidney Fels, and Brian Fisher. Studying IT security professionals: research design and lessons learned. In *CHI 2007 Workshop on Security User Studies: Methodologies and Best Practices*, 2007.

[8] Douglas M Boyle, James F Boyle, and Dana R Hermanson. How to publish in peer-reviewed practitioner accounting journals. *Issues in Accounting Education*, 35(2):19–30, 2020.

[9] Katerina Božič, Alexandre Anatolievich Bachkirov, and Matej Černe. Towards better understanding and narrowing of the science–practice gap: A practitioner-centered approach to management knowledge creation. *European Management Journal*, 40(4):632–644, 2022.

[10] Elizabeth Buie, Susan Dray, Keith Instone, Jhilmil Jain, Gitte Lindgaard, and Arnie Lund. How to bring HCI research and practice closer together. In *Extended Abstracts of the 2010 CHI Conference on Human Factors in Computing Systems*, pages 3181–3184, 2010.

[11] Karoline Busse, Julia Schäfer, and Matthew Smith. Replication: '. . . no one can hack my mind' - revisiting a study on expert and non-expert security practices and advice. In *Proceedings of the Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, pages 117–136, 2019.

[12] Carnegie Mellon University. IoT security and privacy label. https://iotsecurityprivacy.org/, 2024.

[13] Thomas E. Carroll, David Manz, Thomas Edgar, and Frank L. Greitzer. Realizing scientific methods for cyber security. In *Proceedings of the 2012 Workshop on Learning from Authoritative Security Experiment Results*, pages 19–24, 2012.

[14] Parmit K. Chilana, Amy J. Ko, and Jacob Wobbrock. From user-centered to adoption-centered design: a case study of an HCI research innovation becoming a product. In *Proceedings of the 2015 CHI Conference on Human Factors in Computing Systems*, pages 1749–1758, 2015.

[15] Yee-Yin Choong and Mary Theofanos. What 4,500+ people can tell you: employees' attitudes toward organizational password policy do matter. In *Human Aspects of Information Security, Privacy, and Trust: Held as Part of Human-Computer Interaction International 2015, Proceedings 3*, pages 299–310, 2015.

[16] Jacob Cohen. A power primer. *Psychological Bulletin [PsycARTICLES]*, 112(1):155–159, 1992.

[17] Lucas Colusso, Cynthia L. Bennett, Gary Hsieh, and Sean A. Munson. Translational resources: Reducing the gap between academic research and HCI practice. In *Proceedings of the 2017 Conference on Designing Interactive Systems*, pages 957–968, 2017.

[18] Lucas Colusso, Ridley Jones, Sean A. Munson, and Gary Hsieh. A translational science model for HCI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019.

[19] Emma Dahlin. Mind the gap! on the future of AI research. *Humanities and Social Sciences Communications*, 8(1):1–4, 2021.

[20] Peter Dalsgaard and Christian Dindler. Between theory and practice: bridging concepts in HCI research. In *Proceedings of the 2014 CHI Conference on Human Factors in Computing Systems*, pages 1635–1644, 2014.

[21] Jessica Dawson and Robert Thomson. The future cybersecurity workforce: Going beyond technical skills for successful cyber performance. *Frontiers in Psychology*, 9:744, 2018.

[22] Hans deBruijn and Marijn Janssen. Building cybersecurity awareness: The need for evidence-based framing strategies. *Government Information Quarterly*, 34(1):1–7, 2017.

[23] Verner Denvall and Mikael Skillmark. Bridge over troubled water—closing the research–practice gap in social work. *The British Journal of Social Work*, 51(7):2722–2739, 2021.

[24] Gurpreet Dhillon, Kane Smith, and Indika Dissanayaka. Information systems security research agenda: Exploring the gap between research and practice. *The Journal of Strategic Information Systems*, 30(4):101693, 2021.

[25] Kenny Doyle, Zeta Dooly, and Paul Kearney. What's so unique about cyber security? In *Cyber Security and Privacy: 4th Cyber Security and Privacy Innovation Forum, CSP Innovation Forum 2015*, pages 131–139. Springer International Publishing, 2015.

[26] Rachel C. Dreibelbis, Jaclyn Martin, Michael D. Coovert, and David W. Dorsey. The looming cybersecurity crisis and what it means for the practice of industrial and organizational psychology. *Industrial and Organizational Psychology*, 11(2):346–365, 2018.

[27] Josiah Dykstra and Celeste Lyn Paul. Cyber operations stress survey (COSS): Studying fatigue, frustration, and cognitive workload in cybersecurity operations. In *11th USENIX Workshop on Cyber Security Experimentation and Test (CSET 18)*, 2018.

[28] Pardis Emami-Naeini, Janarth Dheenadhayalan, Yuvraj Agarwal, and Lorrie Faith Cranor. An informative security and privacy 'nutrition' label for internet of things devices. *IEEE Security & Privacy*, 20(2):31–39, 2021.

[29] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4):1149–1160, 2009.

[30] Federal Communications Commission. FCC fact sheet - Cybersecurity labeling for internet of things. https://docs.fcc.gov/public/attachments/DOC-400674A1.pdf, 2024.

[31] Simson Garfinkel and Heather Richter Lipford. *Usable security: History, themes, and challenges*. Morgan & Claypool Publishers, 2014.

[32] Gartner. Gartner identifies the top cybersecurity trends for 2023: Security leaders must pivot to a human-centric focus to establish an effective cybersecurity program. https://www.gartner.com/en/newsroom/press-releases/04-12-2023-gartner-identifies-the-top-cybersecurity-trends-for-2023, 2023.

[33] Sabine Geldof and Joannes Vandermeulen. A practitioner's view of human–computer interaction research and practice. *Artifact: Journal of Design Practice*, 1(3):134–134, 2007.

[34] Georgia Tech. Commercialization. https://commercialization.gatech.edu/georgia-tech-research-your-path-commercialization, 2024.

[35] Colin M Gray, Erik Stolterman, and Martin A Siegel. Reprioritizing the relationship between HCI research and practice: bubble-up and trickle-down effects. In *Proceedings of the 2014 Conference on Designing interactive systems*, pages 725–734, 2014.

[36] Matthew Green and Matthew Smith. Developers are not the enemy!: The need for usable security APIs. *IEEE Security & Privacy*, 14(5):40–46, 2016.

[37] Sander Greenland, Stephen J. Senn, Kenneth J. Rothman, John B. Carlin, Charles Poole, Steven N. Goodman, and Douglas G. Altman. Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, 31(4):337–350, 2016.

[38] Jeremy M. Grimshaw, Martin P. Eccles, John N. Lavis, Sophie J. Hill, and Janet E. Squires. Knowledge translation of research findings. *Implementation Science*, 7(1):1–17, 2012.

[39] Marthie Grobler, Raj Gaire, and Surya Nepal. User, usage and usability: Redefining human centric cyber security. *Frontiers in Big Data*, 4:583723, 2021.

[40] Magnus Gulbrandsen and Taran Thune. The effects of non-academic work experience on external interaction and research performance. *Journal of Technology Transfer*, 42:795–813, 2017.

[41] Julie M. Haney, Clyburn Cunningham, and Susanne M. Furman. Towards integrating human-centered cybersecurity research into practice: A practitioner survey. In *Symposium on Usable Security and Privacy (USEC)*, 2024.

[42] Kennenth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2019.

[43] International Computer Science Institute. Usable security and privacy. https://www.icsi.berkeley.edu/icsi/groups/privacy, 2022.

[44] Andrew N. Kadykalo, Rachel T. Buxton, Peter Morrison, Christine M. Anderson, Holly Bickerton, Charles M. Francis, Adam C. Smith, and Lenore Fahrig. Bridging research and practice in conservation. *Conservation Biology*, 35(6):1725–1737, 2021.

[45] Ruogu Kang, Laura Dabbish, Nathaniel Fruchter, and Sara Kiesler. "My data just goes everywhere:" user mental models of the internet and implications for privacy and security. In *Proceeding of the Eleventh Symposium on Usable Privacy and Security (SOUPS 2015)*, 2015.

[46] Hae-Young Kim. Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test. *Restorative Dentistry & Endodontics*, 42(2):152–155, 2017.

[47] Neha Kumar and Nicola Dell. Towards informed practice in HCI for development. *Proceedings of the ACM on Human-Computer Interaction*, 2:1–20, 2018.

[48] Hao-Ping Hank Lee, Lan Gao, Stephanie Yang, Jodi Forlizzi, and Sauvik Das. I don't know if we're doing good. I don't know if we're doing bad': Investigating how practitioners scope, motivate, and conduct privacy work when developing ai products. In *Proceeding of the 33rd USENIX Security Symposium*, 2024.

[49] W. Bentley MacLeod and Miguel Urquiola. Why does the United States have the best research universities? incentives, resources, and virtuous circles. *Journal of Economic Perspectives*, 35(1):185–206, 2021.

[50] Douglas Maughan, David Balenson, Ulf Lindqvist, and Zachary Tudor. Crossing the "valley of death": Transitioning cybersecurity research into practice. *IEEE Security & Privacy*, 11(2):14–23, 2013.

[51] Dale McMorrow. Science of cyber-security. Technical report, The MITRE Corporation, 2010.

[52] Leigh Metcalf and Jonathan Spring. *Using Science in Cybersecurity*. 2021.

[53] Jacqueline Meyer and Giovanni Apruzzese. Cybersecurity in the smart grid: Practitioners' perspective systems (technical report). In *8th Annual Industrial Control Systems Security Workshop*, 2022.

[54] National Cybersecurity Alliance and Cybsafe. Oh behave! The annual cybersecurity attitudes and behaviors report 2023. https://staysafeonline.org/online-safety-privacy-basics/oh-behave/, 2023.

[55] National Institute of Standards and Technology. Special Publication 800-63 Digital identity guidelines. https://pages.nist.gov/800-63-3/, 2017.

[56] National Institute of Standards and Technology. Human-centered cybersecurity. https://csrc.nist.gov/projects/human-centered-cybersecurity, 2023.

[57] Calvin Nobles. Establishing human factors programs to mitigate blind spots in cybersecurity. In *Proceedings of the Fourteenth Midwest Association for Information Systems Conference*, 2019.

[58] Calvin Nobles. Stress, burnout, and security fatigue in cybersecurity: A human factors problem. *Journal of Business and Public Administration*, 13(1):49–72, 2022.

[59] Abinash Panda. Bringing academic and corporate worlds closer: We need pracademics. *Management and Labour Studies*, 39(2):140–159, 2014.

[60] José Paredes, Juan Carlos Teze, Gerardo I. Simari, and Maria Vanina Martinez. On the importance of domain-specific explanations in AI-based cybersecurity systems (technical report). *CoRR*, abs/2108.02006, 2021.

[61] Simon Parkin, Aad Van Moorsel, Philip Inglesant, and M. Angela Sasse. A stealth approach to usable security: Helping IT security managers to identify workable security solutions. In *Proceedings of the 2010 New Security Paradigms Workshop*, pages 33–50, 2010.

[62] Celeste Lyn Paul. Human-centered study of a network operations center: experience report and lessons learned. In *2014 ACM Workshop on Security Information Workers*, pages 39–42, 2014.

[63] Natalie M. Scala, Allison C. Reilly, Paul L. Goethals, and Michel Cukier. Risk and the five hard problems of cybersecurity. *Risk Analysis*, 39(10):2119–2126, 2019.

[64] Raphael Serafini, Marco Gutfleisch, Stefan Albert Horstmann, and Alena Naiakshina. On the recruitment of company developers for security studies: Results from a qualitative interview study. In *Proceedings of the*

Nineteenth Symposium on Usable Privacy and Security (SOUPS 2023), 2023.

[65] Richard Shay, Saranga Komanduri, Patrick Gage Kelley, Pedro Giovanni Leon, Michelle L. Mazurek, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. Encountering stronger password requirements: user attitudes and behaviors. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, pages 1–20, 2010.

[66] C. Estelle Smith, Eduardo Nevarez, and Haiyi Zhu. Disseminating research news in HCI: Perceived hazards, how-tos, and opportunities for innovation. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[67] Jeremiah D. Still. Cybersecurity needs you! *Interactions*, 23(3):54–58, 2016.

[68] Alexander Styhre. The influence of neoliberalism and its absence from management research. *International Journal of Organizational Analysis*, 22(3):278–300, 2014.

[69] Gail M. Sullivan and Richard Feinn. Using effect size—or why the p value is not enough. *Journal of Graduate Medical Education*, 4(3):279–282, 2012.

[70] Mary Theofanos. Is usable security an oxymoron? *Computer*, 53(2):71–74, 2020.

[71] Mary Theofanos, Simson Garfinkel, and Yee-Yin Choong. Secure and usable enterprise authentication: Lessons from the field. *IEEE Security & Privacy*, 14(5):14–21, 2016.

[72] University of Maryland College of Information Studies. Sociotechnical Cybersecurity (STC) Interest Group. https://ischool.umd.edu/centers-and-labs/stc/, 2023.

[73] Michael Veale, Max Van Kleek, and Reuben Binns. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2018.

[74] Verizon. 2023 data breach investigations report. https://www.verizon.com/business/resources/reports/dbir, 2023.

[75] Ryan West, Christopher Mayhorn, Jefferson Hardee, and Jeremy Mendel. The weakest link: A psychological perspective on why users make poor security decisions. In *Social and Human Elements of Information Security: Emerging Trends and Countermeasures*, pages 43–60, 2009.

[76] Robin Whittemore and Gail Melkus. Design decisions in research. *e-Source Behavioral Social Sciences Research*, n.d.

[77] Mary Ellen Zurko and Julie Haney. Usable security and privacy for security and privacy workers. *IEEE Security & Privacy*, 21(1):8–10, 2023.

APPENDIX A: SURVEY INSTRUMENT

Terminology

Security *will be used as shorthand for cybersecurity.*

Human-centered security *(sometimes called "usable security") considers the human, social, and organizational factors related to security processes, technologies, products, policies, etc. It involves the relationships and interactions between people and cybersecurity, including people's perceptions, the challenges they encounter, and designing usable systems, products, and services that also result in improved security outcomes.*

Research *refers to human-centered security research you are currently conducting or have conducted in the past.*

Practitioners *are individuals who engage in a profession either directly related to security or significantly involving security considerations. Examples include, but are not limited to: Security practitioners – for example, administrators, analysts, architects, consultants, trainers whose primary job involves security IT practitioners - for example, system administrators, help desk staff, system and network architects Developers – for example, software and hardware developers who implement security features or mechanisms in their products Organizational leadership – for example, managers and executives Policy makers who include security considerations in their directives Educators and trainers who teach people about security.*

Practitioner resources *are those sources that are developed by or written for practitioners and are not published in research forums. Examples include industry reports and market surveys; technical white papers, standards, and guidelines; regulations and policies; and government reports.*

Information About You and Your Research

1) What is your current research position? If you are also a practitioner, you will have an opportunity to indicate that in the next question.

- ☐ Undergraduate student
- ☐ Graduate student
- ☐ Tenure-track faculty
- ☐ Non-tenure-track faculty
- ☐ Other type of researcher (non-faculty)
- ☐ Other (please specify)

2) Have you ever worked as a software/hardware developer, a security practitioner, or an IT practitioner?

- ○ Yes, in the past
- ○ Yes, and I currently still am a practitioner

○ No

3) What kind of practitioner have you been? Select all that apply.
(Only asked if "Yes, in the past" or "Yes, and I currently still am a practitioner" was selected in Question 2)

☐ Security practitioner

☐ IT practitioner

☐ Software or hardware developer

☐ Manager or executive

☐ Policy maker

☐ Educator/trainer

☐ Other (please specify)

4) How many years have you conducted human-centered security research?

○ Less than 1

○ 1-5

○ 6-10

○ 11-15

○ 16-20

○ More than 20 years

5) Which of the following best describes your current, primary organization/institution?

○ Academic

○ Private industry

○ Non-profit

○ Government

○ Other (please specify)

6) In which region is your current organization?

○ Africa

○ Asia

○ Europe

○ North America

○ Oceania

○ South America

○ Caribbean Islands

○ Pacific Islands

7) Which type of funding has supported your human-centered security research? Select all that apply.

☐ Public funding from a government (international, national, or local) or other organization supported in part or in full by revenue generated by a government

☐ Private funding from a corporate organization or other organization not publicly funded

☐ Private funding from a corporate organization or other organization not publicly funded

☐ No specific funding

☐ I'm not sure

☐ Other (please specify)

8) What user populations have been the focus of your research? Select all that apply.

☐ General public end users

☐ Organizational end users (employees)

☐ Security practitioners

☐ Students

☐ IT practitioners

☐ Developers

☐ Organizational leadership

☐ Policy makers

☐ Educators/trainers

☐ Other (please specify)

9) Which populations could make use of or put into practice the implications and recommendations from your human-centered security research? Select all that apply.

☐ General public end users

☐ Organizational end users (employees)

☐ Students

☐ Security practitioners

☐ IT practitioners

☐ Developers

☐ Organizational leadership

☐ Policy makers

☐ Educators/trainers

☐ Other (please specify)

10) How often does your human-centered security research directly impact security practice?
Never - Rarely - Sometimes - Often - Always - Don't Know

Research Conceptualization

Remember: For the purposes of this survey, research refers to human-centered security research you are currently conducting or have conducted in the past. Practitioner resources are those sources that are developed by or written for practitioners and are not published in research forums (e.g., industry reports and market surveys; technical white papers, standards, and guidelines; regulations and policies; and government reports).

11) How often do you consult practitioners or practitioner resources when performing the following research activities?
Never - Rarely - Sometimes - Often - Always

Identifying a new research topic or problem

Developing research questions or hypotheses

Conducting a literature review

12) What do you think is the level of importance of consulting practitioners or practitioner resources when performing the following research activities?
Not Important - Slightly Important - Somewhat Important - Moderately Important - Extremely Important

Identifying a new research topic or problem

Developing research questions or hypotheses

Conducting a literature review

13) What is the level of challenge you have experienced when consulting practitioners or practitioner resources for the following research activities?
Not Challenging - Slightly Challenging - Somewhat Challenging - Moderately Challenging - Extremely Challenging - No Experience to Judge

Identifying a new research topic or problem

Developing research questions or hypotheses

Conducting a literature review

Study Design

14) How often do you consult practitioners or practitioner resources when performing the following study design activities for your human-centered security research?
Never - Rarely - Sometimes - Often - Always

Deciding which research methodology is most appropriate

Developing and piloting research instruments (such as surveys and interview protocols) and experiments

Developing a plan for sampling/recruiting research participants

15) What do you think is the level of importance of consulting practitioners or practitioner resources when performing the following study design activities?
Not Important - Slightly Important - Somewhat Important - Moderately Important - Extremely Important

Deciding which research methodology is most appropriate

Developing and piloting research instruments (such as surveys and interview protocols) and experiments

Developing a plan for sampling/recruiting research participants

16) What is the level of challenge you have experienced when consulting practitioners or practitioner resources for the following study design activities?
Not Challenging - Slightly Challenging - Somewhat Challenging - Moderately Challenging - Extremely Challenging - No Experience to Judge

Deciding which research methodology is most appropriate

Developing and piloting research instruments (such as surveys and interview protocols) and experiments

Developing a plan for sampling/recruiting research participants

Participant Recruitment

17) Have you conducted research for which you recruited practitioners as participants? Select all that apply.

☐ Yes, for surveys

☐ Yes, for interviews

☐ Yes, for focus groups or workshops

☐ Yes, for an experiment

☐ Yes, for another purpose

☐ No

18) In what ways have you attempted to recruit practitioners? Select all that apply. *(Only asked if "No" was NOT selected in Question 17)*

☐ Professional contacts

☐ Snowballing

☐ Online forums

☐ Mailing lists

☐ Social media (for example, Twitter, Instagram, Reddit, Facebook)

☐ Flyers

☐ Online advertisements (for example, Craigslist)

☐ Research panels or crowdsourcing platforms (for example, Mechanical Turk, Prolific, Qualtrics?)

☐ Other (please specify)

19) What is the level of challenge you have experienced when recruiting practitioners for your research? *(Only asked if "No" was NOT selected in Question 17)*
Not Challenging - Slightly Challenging - Somewhat Challenging - Moderately Challenging - Extremely Challenging

Data Analysis

20) How often do you consult practitioners or practitioner resources when performing the following data analysis activities for your human-centered security research?
Never - Rarely - Sometimes - Often - Always

Analyzing data (for example, statistical analysis or qualitative data coding)

Developing implications, recommendations, or solutions based on research results

21) What do you think is the level of importance of consulting practitioners or practitioner resources when performing the following data analysis activities?
Not Important - Slightly Important - Somewhat Important - Moderately Important - Extremely Important

Analyzing data (for example, statistical analysis or qualitative data coding)

Developing implications, recommendations, or solutions based on research results

22) What is the level of challenge you have experienced when consulting practitioners or practitioner resources for the following data analysis activities?
Not Challenging - Slightly Challenging - Somewhat Challenging - Moderately Challenging - Extremely Challenging - No Experience to Judge

Analyzing data (for example, statistical analysis or qualitative data coding)

Developing implications, recommendations, or solutions based on research results

23) Thinking about your research conceptualization, recruitment, design, and analysis activities, what barriers, if any, have you encountered when attempting to consult practitioners or practitioner resources? Select all that apply.

- ☐ There is little or no incentive for me to consult these.
- ☐ My research is not relevant to practitioners.
- ☐ Practitioner problems aren't of interest to my funding sources.
- ☐ There is little funding or resources to do this.
- ☐ I don't have time.
- ☐ Practitioners don't have time to participate.
- ☐ Practitioners don't see the value in participating.
- ☐ Practitioners' organizations don't permit them to participate.
- ☐ I'm not sure how to best reach practitioners.
- ☐ Practitioner resources may not be based on rigorously gathered evidence.
- ☐ Practitioners don't have a research background, so their help would be limited.
- ☐ I'm not sure, but I've had problems.
- ☐ I haven't experienced any barriers, even though I've consulted practitioners and practitioner resources.
- ☐ I haven't experienced any barriers because I haven't tried to consult practitioners or practitioner resources.
- ☐ Other (please specify)

Research Dissemination

24) How often are your research outputs (e.g., papers/articles, presentations, blogs, tools) targeted at practitioners?
Never - Rarely - Sometimes - Often - Always - I do not produce or have not yet produced research outputs

25) In what ways have you disseminated your practitioner-focused research outputs? Select all that apply. *(Only asked if "Never" and "I do not produce or have not yet produced research outputs" were NOT selected in Question 24)*

- ☐ Discussions with practitioners
- ☐ Papers/articles in practitioner-focused publications
- ☐ Presentations at practitioner-focused conferences, meetings, or other events
- ☐ Podcasts
- ☐ News media
- ☐ Videos
- ☐ Websites, blogs, other online forums
- ☐ Social media
- ☐ Mailing lists
- ☐ Tools or other software or hardware
- ☐ Knowledge and data repositories
- ☐ Government publications

- ☐ Standard documents
- ☐ Other (please specify)

26) What do you think is the level of importance of producing or contributing to research outputs targeted at practitioners?
Not Important - Slightly Important - Somewhat Important - Moderately Important - Extremely Important

27) In your opinion, what is the extent to which practitioners would be interested in having research outputs shared with them?
Not Interested at All - Slightly Interested - Somewhat Interested - Moderately Interested - Extremely Interested

28) What is the level of challenge you have experienced when producing or contributing to research outputs targeted at practitioners?

Not Challenging - Slightly Challenging - Somewhat Challenging - Moderately Challenging - Extremely Challenging - No Experience to Judge

29) What barriers, if any, do you encounter when producing or contributing to research outputs targeted at practitioners? Select all that apply.

- ☐ There is little or no incentive for me develop research outputs for practitioners.
- ☐ There is little funding or resources to do this.
- ☐ I don't have time.
- ☐ I am concerned that my research will be misinterpreted.
- ☐ I'm not sure how to translate research topics into content valuable to practitioners.
- ☐ I'm not sure where to disseminate my research results.
- ☐ It is difficult to get my article/presentation accepted to practitioner-focused publications and forums.
- ☐ Lack of interest or uptake from practitioners
- ☐ I haven't experienced any barriers.
- ☐ I have not attempted to report results to practitioners.
- ☐ Other (please specify)

30) Please share any other thoughts you have regarding interactions with practitioners in human-centered security research.

———

APPENDIX B: CONFERENCES USED FOR RECRUITMENT

- Symposium on Usable Privacy and Security (SOUPS) 2020 – 2022
- IEEE Symposium on Security and Privacy 2020-2022
- USENIX Security Symposium 2020 – 2022
- ACM Conference on Human Factors in Computing Systems (CHI) 2020 - 2023
- Symposium on Usable Security (USEC) 2021 – 2022
- European Workshop on Usable Security (EuroUSEC) 2020 - 2022
- AsiaUSEC 2020

- Socio-technical Aspects in Security Workshop 2020 – 2021
- Human-Computer Interaction for Cybersecurity, Privacy, and Trust (affiliated conference at International Conference on

Human-Computer Interaction) 2020 - 2022

- Human Aspects of Information Security and Assurance (HAISA) 2020 - 2022

Comparing Teacher and Creator Perspectives on the Design of Cybersecurity and Privacy Educational Resources

Joy McLeod
Carleton University

Leah Zhang-Kennedy
University of Waterloo

Elizabeth Stobert
Carleton University

Abstract

Various educational resources have been developed to teach children about cybersecurity and privacy. Our qualitative interview study with 15 middle school teachers and 8 creators of cybersecurity educational resources compares and analyzes the design considerations of cybersecurity resource creators with the resource selection strategies and classroom practices of teachers in their delivery of cybersecurity lessons to middle school students. Our thematic analysis showed that teachers predominately used free, low-tech, modular, and modifiable resources such as lesson plans, short educational videos, and segmented learning modules to fit their classroom teaching needs. The topics focus on helping students develop critical thinking skills rather than technical knowledge. Creators, on the other hand, focused their resource design considerations primarily on cybersecurity trends and students' media learning preferences, such as developing games and other types of interactive content to increase engagement. We highlight areas of misalignment between creators' design considerations compared to how teachers access and deliver cybersecurity and privacy lessons to students.

1 Introduction

Cybersecurity and privacy have emerged as a topic of concern for parents, educators, and policymakers [11] as people are using an ever-expanding number of services to live and work, and the importance of knowing how to stay safe online, protect personal information and verify the authenticity of information found online has never been greater [6, 16, 17, 23].

Due to the high potential for exposure to online risks, a focal point of intervention has been the development of initiatives that aim to educate young people about online risks. The goal is for young people to develop their knowledge about cybersecurity and privacy so they can critically examine their online experiences and protect themselves online. Teachers are increasingly asked to assume the responsibility of educating young people to thrive as digital citizens and future employees [16–18]. However, teachers may not be properly equipped with their own knowledge of security and privacy to teach these subjects to their students [4, 8, 16, 18, 21, 31, 33].

Various cybersecurity education resources for the K-12 classroom [1, 14, 26] have been created to help teachers carry out this important task. Previous research [39] found that about half of the tools and resources in the last decade are aimed at children and youth. However, there is limited understanding of how teachers utilize these resources in the classroom [23], making it difficult to assess how effectively these resources meet the needs of teachers and students.

This paper aims to compare the teaching practices of middle school teachers with the design considerations of creators of cybersecurity educational resources. Our goal is to determine if the process of creating and distributing resources by content creators aligns with how teachers discover and use these resources in the classroom. This intersection between creators and teachers in cybersecurity education has not been explored before. We define a resource creator (hereby referred to as "creators") as a stakeholder who has contributed to the design of cybersecurity educational materials. A creator could be a designer, developer, researcher, or project manager who has experience in industry or academia creating cybersecurity educational resources. Our research questions are:

RQ1 What do teachers consider when choosing cybersecurity and privacy educational resources to use in the classroom and how do they assess learning outcomes?

RQ2 What do creators consider when curating, designing, and evaluating cybersecurity and privacy resources for use in the classroom?

RQ3 How well do creators' design considerations and processes for the educational content and format of delivery align with the needs of teachers and students to teach and learn about cybersecurity and privacy?

To answer our research questions, we interviewed 15 middle school teachers who have taught tweens (aged 10–13 years) and 8 resource creators to understand their processes, challenges, experiences, and needs. We focus on middle school teachers because their tween students are a vulnerable demographic that needs significant support and guidance from teachers as they navigate digital media [10, 24].

We analyzed our data using thematic analysis and found that teachers were predominantly using freely available, low-tech, lesson-oriented resources in their teaching, such as lesson plans, short videos, and segmented learning modules, and generally found these resources effective. Their considerations in choosing resources focused on alignment with their classroom teaching needs and how well the resources supported inquiry and critical thinking skills. Most taught cybersecurity and privacy as an ad hoc reaction to classroom incidents, such as cyberbullying, which influenced their preferences for finding and choosing resources. Teachers reported a variety of assessment methods to measure learning outcomes, but showed a preference for critical reflection over formal assessment due to the sensitivity of the topics.

Creators showed a general awareness of the time constraints of teachers related to curricular expectations and the technical challenges teachers face in incorporating cybersecurity resources into the classroom. However, they prioritized the needs and learning preferences of the primary target audience of the educational resource, such as design considerations that make the resources engaging and fun for young people. Furthermore, our investigations into creators' design processes show that they lack centralized guidance on what baseline topics should be taught, causing them to develop resources based on current cybersecurity trends and funding opportunities.

2 Background and Related Work

Government, not-for-profit organizations, and academic researchers make available a variety of resources to assist teachers in teaching topics of privacy, cybersecurity and digital literacy to their students. Resources are generally provided online and organized by the curricular expectations, geography, topic, grade, and media type [12, 15, 26, 27, 35]. In more structured programs, the lessons are organized predominantly by topic and grade in discrete packages [3, 30, 34], such as Google's Applied Digital Skills curriculum on digital footprints, online scams, cyberbullying, and more [13].

Supporting resources for teachers are often included with the educational tools as lesson and facilitation guides to help them use the resources and deliver the lesson. Other related teacher resources include materials such as slides, tip sheets, videos, printable classroom activities, quizzes, and assessments [12, 15, 26, 27, 30, 34, 35].

2.1 Cybersecurity Educational Tools and Resources

A variety of multimedia tools such as games, videos, tabletop games, learning modules, and comics [37, 38, 40] have been developed to teach people of all ages about cybersecurity [39]. Games, in particular, are the most popular type of resource, as they are believed to be a particularly powerful experiential learning tool [23, 39].

In a systematic review of multimedia tools for cybersecurity awareness and education created between 2000 and 2019, Zhang-Kennedy and Chiasson [39] identified that approximately 43% of the tools are tailored to children and youth, but most tools lack evaluations to support the effectiveness of the learning outcomes. Another systematic review of the literature on children's cybersecurity awareness in 2021 [32] added to this by pointing out the lack of valid evaluation methods, theoretical frameworks, small sample sizes, and a bias toward early signs of positive results.

Although educational and training resources created to improve the general public's cybersecurity and digital literacy could be used by teachers(e.g., Cybersec101 [3]), public professional development training resources tailored to teachers are rare and focus primarily on students' privacy. For example, iKeepSafe [19] has an educator training course on data privacy in education. Common Sense Education [9] offers free teacher privacy compliance training to protect student privacy. The Student Privacy Compass [36] has a series of student privacy training for educators that touches on a variety of topics, including training on why students need to learn about privacy and the key topics to teach.

2.2 Challenges in Teaching Cybersecurity and Digital Literacy

Few studies have explored how teachers are currently using tools and resources to teach cybersecurity and digital literacy, the challenges they face, and their perceptions of students' skills and competencies.

Weinstein et al. [20] surveyed K–12 teachers in the U.S. and found that approximately 60% used some type of digital literacy curriculum or resource with students in the classroom. Furthermore, 70% of teachers reported teaching at least one type of digital literacy competency, with the most common being cyberbullying (46%) and privacy and safety (44%).

Maqsood and Chiasson [24] conducted a study with 21 Canadian elementary school teachers to understand the risks teachers were seeing their 10 to 13 year old students. They found that teachers regularly helped their students mitigate risks from minor policy violations to more serious forms of

cyberbullying. However, teacher reported a lack of knowledge, training, and support to address issues at their schools.

Corradini and Nardelli [10] conducted a study with 2,229 Italian primary and secondary school teachers' about their perceptions of their students' digital awareness. They found that teachers felt students should be better prepared to recognize risks when using digital technologies, pay more attention to protecting their personal data and privacy, and learn media literacy in terms of measuring the reliability of news on social media. Similar to the findings of Maqsood and Chiasson [24], the Italian teachers also reported that they needed additional training to improve their own digital awareness and administrative support in their activities.

Kumar et al. [22] conducted focus groups with 25 educators to better understand what privacy and security meant to them. They found that technology use is an integral part of the elementary school classroom and that educators consider digital privacy and security through the lens of their curricular and classroom management goals.

Nicholson et al. [28] conducted a study with 50 secondary school children aged 12-14 and found that teachers described the education process as a "*piecemeal approach*," with students reporting learning about related and non-technical aspects of privacy and security (e.g., cyberbullying) through sporadic lessons and not in a consistent, ongoing way.

Martin et al. [25] conducted a study with 107 K-12 educators to understand their perceptions of their students' digital citizenship knowledge and practices. They found that educators who taught digital citizenship had higher perceptions of their students' digital citizenship practices than other educators. Teachers reported the need for more training, resources, and activities relating real-world examples, and integrating digital citizenship into curriculum.

2.3 Research Gap

Significant work has been done to develop privacy and cybersecurity educational materials for children. However, there is a lack of studies that focus on teachers' perspectives when teaching these topics [22, 24]. While there are some studies that aim to evaluate specific resources, none of these studies explores how teachers approach these subjects with their students. Our goal is to compare teachers and creators' perspectives on teaching cybersecurity and privacy, to identify whether these materials are being designed well, accessed widely, and used effectively.

In our work, we interviewed 15 teachers and 8 creators to compare their perspectives on cybersecurity and privacy education, and identified overlaps and divergences between teachers' and creators' perspectives. Based on our findings, we highlight areas of misalignment between creators' design considerations compared to how teachers access and deliver cybersecurity and privacy lessons to students.

3 Methodology

We conducted semi-structured interviews with teachers and resource creators. We interviewed 15 pre-secondary school teachers and 8 creators. Both studies followed the same basic methodology and received clearance from our institution's Research Ethics Board.

3.1 Procedure

Study participants completed a brief screening questionnaire before being invited to participate in an online interview lasting 60 to 75 minutes. The interviews were audio-recorded and transcribed using Trint[1] and manually checked for accuracy. The participants were remunerated $45 CAD.

The teachers' pre-interview questionnaire (see Appendix 9) asked demographic questions, as well as questions about teachers' experience with cybersecurity and privacy topics and the resources they use. The teacher interview questions (see Appendix 11) explored the following areas:

- *Practices:* How do teachers teach cybersecurity and privacy to their students?

- *Selection:* How do teachers find and choose the resources they use to teach cybersecurity and privacy?

- *Effectiveness:* How effective do teachers find these resources?

- *Experience:* What do teachers like and dislike about these resources?

To ground teachers' responses in their classroom experiences, teachers participating in the interview were asked to bring examples of resources they had previously used to teach cybersecurity or privacy, and to explain how and why they were used.

The creators' pre-interview questionnaire (Appendix 10) asked demographic questions, and about creators' experiences designing educational materials for teaching cybersecurity and privacy, and what topics and issues they considered in the design of these materials. The creator interviews (Appendix 12) were structured around the following topics:

- *Processes:* How do creators go about developing educational resources for cybersecurity and privacy in their organizations?

- *Dissemination:* How do creators make schools and teachers aware of these resources?

- *Improvement:* How could creators' design processes or resources be improved?

- *Strategies:* What strategies do creators use when designing resources for different age groups?

---

[1] https://trint.com/

3.2 Participants

We recruited participants for both studies using a combination of snowball sampling, social media, and emails.

3.2.1 Teachers

To qualify for the study, teachers had to be Canadian and have had experience teaching cybersecurity and privacy to pre-secondary school students in the last two years. We limited recruitment to Canadian teachers so they could share experience in a similar educational system. Teacher recruitment notices were emailed to local contacts, teacher-oriented associations, and school mailing lists (with the approval of school boards). We also posted recruitment notices to relevant Facebook and Reddit groups.

In total, we interviewed 15 teachers from 11 schools in three of the largest Canadian provinces[2]. Table 1 summarizes the demographics of the teachers. The majority (67%) were female, and the remainder (33%) were male. Our participants had a wide range of teaching experience from 1 to 35 years ($Mdn = 15$). More than half (53%) were mid-career professionals over the age of 40. All had experience teaching middle grades, though many also had experience teaching a broader range of students ranging from kindergarten to grade nine. All but one participant (93%) taught in public schools. The majority of the teachers (87%) had an educational background in arts, languages, or education, with only one having a background in science.

3.2.2 Creators

We broadly defined a creator as a stakeholder who has professional experience in creating cybersecurity educational resources. As we did not limit their roles to the implementation of resources, these individuals could include designers, developers, researchers, project managers, and educational directors. As a starting point, the lead researcher emailed researchers and practitioners listed in the Canadian Cybersecurity Awareness Stakeholders Teleconference Report [2] and asked those contacts to pass the recruitment notice along to their contacts. We were able to recruit eight creators, summarized in Table 2. Of these eight, half were female. The majority (88%) were based in Canada, and one participant (C8) was based in the United States. Six participants (75%) were mid- to late-career professionals 40 years or older, with two over 60 years of age.

In total, our creator participants represented eight different organizations that represented the not-for-profit, public, and private sectors. We do not suggest that our sample is representative of creators in cybersecurity education. However, our sample includes creators with various educational work experiences. Three of the participants (38%) had been creating

cybersecurity and privacy resources for 10+ years, and the remaining five participants (63%) had 5–9 years of experience. More than half (63%) of the participants reported being in senior leadership positions; the other three reported positions related to cybersecurity education research and consulting.

In terms of the educational levels of the participants, two (26%) had bachelor's degrees, three had master's degrees (37%), and three had doctoral degrees (37%). Six participants (75%) reported that their education was directly related to their work creating resources related to privacy and education, and the other two (25%) reported that although their education was not focused on these areas, they had learned the skills and knowledge they needed on the job.

3.3 Reflexive Thematic Analysis

We used reflexive thematic analysis [5,7] for our qualitative analyses in both studies. This approach emphasizes the researcher's active and reflexive role in knowledge production, and acknowledges that codes are understood to represent the researcher's interpretation of meaning and patterns within the data set [7]. The lead researcher had some elementary school teaching experience and conducted all interviews. They were most closely involved with the research, giving them the most relevant contextual experience for the analysis. While codebooks were developed as part of the analysis process for both studies, coding reliability was not calculated due to the reflexive nature of data coding [7]. Instead, intermediate results were regularly reviewed and discussed with two other researchers to help refine the coding categories and extract meaning from the data.

The first stage of our thematic analysis was coding. The lead researcher familiarized themselves with the data by reading and re-reading the transcripts and adding annotations and comments line-by-line using Microsoft Word's commenting feature. This initial process focused on noting key terms and the underlying idea of each response to help get a sense of emergent patterns in the data. Once this was completed, the lead researcher began the process of assigning preliminary codes [5]. The process was repeated for each study.

For the teacher study, we coded 273 pages of transcriptions generated from over 21 hours of audio recordings of interviews. In total, we created 230 codes. For the creator study, we coded 124 pages of transcriptions generated from over 8.5 hours of audio recordings of interviews. In total, we created 280 open codes.

Following open coding, we transitioned to the process of identifying themes. Using Miro[3], we examined our open codes for the underlying patterns. We organized the uncategorized open codes into themes [5], which are presented below in Sections 4 and 5. We attribute direct quotes by appending the letter "T" (e.g., T4) or "C" (e.g., C8) to identify the participant as either a teacher or a creator.

---

[2]Canada's four largest and most populous provinces are Ontario, Quebec, British Columbia, and Alberta.

[3]Miro: https://miro.com

Table 1: Teacher demographics.

ID	Gender	Age	Educational Background	Exp. (years)*	Grades	School	Province
T1	Female	30–39	Drama, English (Minor)	8	7-12	Public	Ontario
T2	Female	30–39	Criminology	7	5-12	Public	Ontario
T3	Male	20–29	Arts, French, Education, History (Minor)	1	5-6	Public	Quebec
T4	Male	30–39	Drama, History	8	6	Public	Ontario
T5	Female	40–49	English Lit., Child Psychology (Minor)	15	5–8	Public	Ontario
T6	Male	50–59	History, Fine Art, Music	27	K–11	Private	Ontario
T7	Male	30–39	*Unspecified*	12	7–11	Public	Quebec
T8	Female	40–49	Kinesiology	20	6	Public	Ontario
T9	Female	20–29	Development Studies, English, Education	4	6	Public	Quebec
T10	Female	50–59	History, Classical Studies	31	3–6	Public	Ontario
T11	Female	60+	Education	35	1–12	Public	Alberta
T12	Female	50–59	History	31	7–8	Public	Ontario
T13	Female	50–59	History, English	27	7–8	Public	Ontario
T14	Female	30–39	Education	5	6	Public	Alberta
T15	Male	40–49	Arts, Education, Social Studies (Minor)	19	7	Public	Alberta

*Years of work experience related to general teaching

Table 2: Creator demographics.

ID	Gender	Age	Educational Background	Highest Degree	Exp. (years)*	Type of Organization	Organization Size	Job Title
C1	Male	40–49	Theatre, English, Education	Bachelor's	13	Not-for-profit	10–49	Director of Education
C2	Male	40–49	Info. Systems, Bus. Mgmt., Criminology	Doctoral	7	Public sector	0–9	Executive Director
C3	Female	40–49	*Unspecified*	Bachelor's	10	Public sector	1000–4999	Supervisor
C4	Male	60+	Engineering, Bus. Admin., Education	Doctoral	5	Both sectors	100–499	President
C5	Female	20–29	Public Policy	Masters	7	Public sector	10–49	Senior Manager
C6	Female	30–39	Computer Science, HCI, Usable security	Doctoral	7	Public sector	1000–4999	Post-doctoral Fellow
C7	Male	60+	Biochemistry, Education	Masters	5	Public sector	0–9	Educational Consultant
C8	Female	40–49	Linguistics	Masters	9	Not-for-profit	10–49	Research Scientist

*Years of work experience creating cybersecurity educational resources.

4 Teachers' Perspectives

Figure 1 shows commonly reported topics taught to students, including "*Cyberbullying*" (87%), "*Cybersecurity*" (80%), and "*Privacy*" (73%). The least commonly taught subjects were "*Authentication*" (20%), "*Gambling*" (20%), and *Pornography*" (20%).

The three most popular resource types used by teachers in our study were lesson plans (87%), learning modules (67%), and live videos (47%). The least-used resource types were comics and gamified activities (13%), and none of our participants reported ever using non-digital or mobile games. Resources that were frequently mentioned were from Media Smarts, Common Sense Media, and Teachers Pay Teachers.

4.1 Resource Discovery

The majority of teachers reported relying on Google searches using key terms and the grade level, highlighting the importance of search engine optimization to improve the chances of teachers finding relevant resources. More experienced teachers reuse the resources they have accumulated over time, and others go directly to trusted organizations' websites (e.g., MediaSmarts, Common Sense Media), or eliciting recommen-

dations from trusted colleagues.

Our teacher participants reported that it is uncommon for their school boards or Ministries of Education to provide curriculum teaching resources on cybersecurity and digital literacy. While elements of these topics are taught as part of the health science and media literacy curriculum, most teachers reported that due to competing curricular priorities, they addressed these topics sporadically or only after a negative event occurred at school. For example, T15 commented, *"By and large, it's only brought up outside of health class when someone gets in trouble. Like, it's not something that is generally talked about in a regular, neutral fashion."* Only a few participants said they take a proactive approach, such as dedicating a week to an entire program, such as the suite of lessons developed by Common Sense Media. This suggests that there is considerable variability in how and when teachers address these topics. For the most part, teachers reported approaching cybersecurity and privacy topics reactively and ad hoc.

4.2 Resource Selection

Teachers had a myriad of considerations when choosing between resources. Their main concerns were how well the resource met their own needs while balancing that against
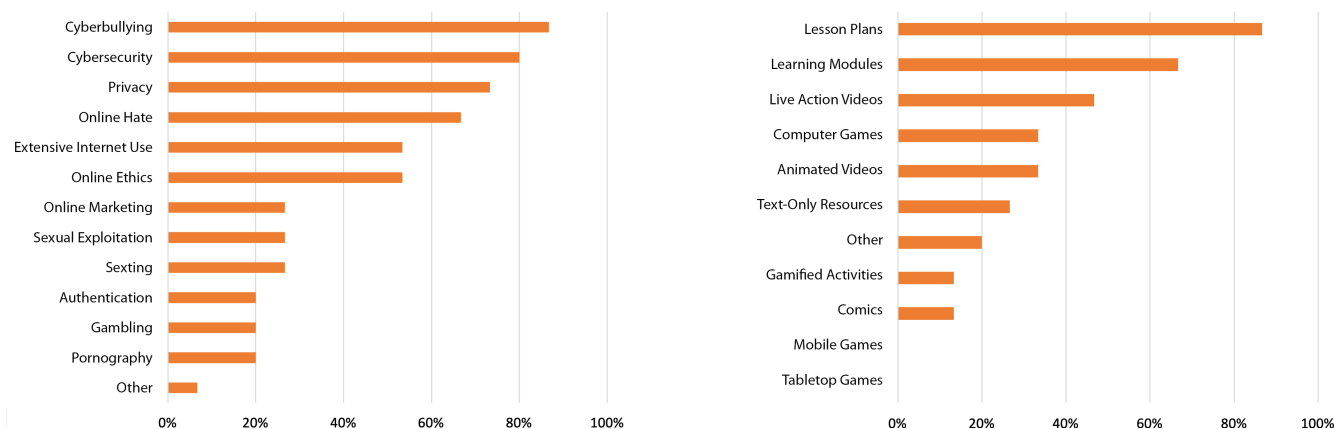
Figure 1: Percentage of teachers who had taught various cybersecurity topics (left) and used various types of resources (right).

how appealing and relatable the resource would be to their students. Although our participants reported using a variety of resources, we found that, in general, simple and accessible resources such as lesson plans, short videos and animations, computer games, quizzes, and classroom activities were the preferred resources used to teach cybersecurity and privacy compared to mobile games, board games, or comics.

4.2.1 Critical Thinking

Teachers preferred resources that promote inquiry and critical thinking skills. This was best done by providing "minds-on" (T12) questions, discussion prompts, or challenges where students were encouraged to investigate things within the resource. For example, T15 shared a resource by CIVIX, a Canadian not-for-profit that they felt did this well.

> And the way that the CTRL-F program is designed, it starts with a question of some kind. And then they have to try and go into, well, how exactly does this work? And there's a lot of critical thinking for them to go back and rethink stuff that they've been assuming about their own practices and the Internet in general for a long time.

Teachers frequently used discussions and reflection questions. T8 highlighted that they thought that discussion was the best methodology to engage students: *"It fuels the active exchange between the students. And I think it's actually a pretty good way to teach those subjects to get them engaged, to get them to share what they think. . . and feel kind of comfortable asking about these things."*

Interestingly, the sentiment was that discussion was somehow *"not about school. . . this isn't about learning,"* (T9) or that these discussions would be something that students would respond to differently because they would not see it as a traditional part of their education. This highlights an interesting tension as it suggests that both teachers and their students may frame approaches focused on rote learning and grading as potentially undermining the goal of the lesson. Teachers re-

ported using resources that incorporated stories, role-playing, and scenarios to help their students imagine potential scenarios and how they would respond to them as a means to help students learn about these topics.

4.2.2 "Safe" Topics

Teachers emphasized the importance of making students feel safe in the classroom. As such, they took great pains to create a sense of psychological safety when discussing sensitive topics around cybersecurity and privacy topics. Students may be uncomfortable because these topics are taboo in their households or because they have fears of being judged for their own behaviour. Due to the potentially difficult and in some cases taboo nature of some topics in cybersecurity and privacy, some teachers expressed concern that covering these topics put them at risk of overstepping their professional boundaries, which might result in professional reprisals. An example was the risk of being listed in the "blue pages," a disciplinary mechanic of the College of Teachers where teachers found to be incompetent or guilty of professional misconduct are publicly listed [29].

> I feel that teachers are not given the full freedom to really provide their best because we are so damn scared of showing our name in the blue pages. . . with all good intentions I tried to teach all these things that I am teaching now which were not very well accepted eight or nine years ago. . . I don't feel comfortable talking about it. (T5)

4.2.3 Relevant and Relatable

The perceived relevance of a resource was of paramount concern for teachers, as they noted their students are quick to reject resources that do not relate to their current interests and experiences. As a result, teachers highlighted the importance of keeping resources up-to-date both in terms of content and

physical design, including well-known stories and technologies relevant to their students' experiences. Further, it should include timely stories and situations relevant to their own experience or local community. T14 shared an experience that highlights this sensitivity:

> *The kids are always very quickly moving on to the next big thing that everybody's using. And I think staying on top of what that is and reflecting that in the resources is really important because we had some group come to do a talk on boundaries and stuff like that and they had Fortnite as one of their slides and all the kids just go up and, you know, Fortnite sucks and blah blah blah. So making sure that it stays relevant to what they're interested in it...*

Multiple teachers reported that they had modernized a resource themselves by changing a referenced technology or game to a more relevant example (e.g., changing a Facebook post to a TikTok post), or by finding widely known news stories, memes, and pop culture references currently popular on social media sites to help build interest and engagement with their students. Teachers also reported looking for resources that had a local focus where possible, whether to discuss a topic that was particularly relevant to their community, or something that they thought their students in particular needed to be aware of.

Teachers were concerned about how relatable the subject matter of the material was to their students. As such, they gravitated to resources that provided a clear rationale for why the lesson is important and how it relates to the experiences that their students have had. Teachers reported that they searched for stories from other young people who had experience with the topic to help communicate the importance of the topic and make it more relatable to their students. We also found instances where teachers gravitated towards resources that included information that their students would find shocking or interesting. As such, some teachers reported looking for resources that referenced highly publicized news stories or resources that incorporated real-life examples. For example, teachers using videos that had information shared by other children around their students' age, which they felt made it especially engaging for their students because it "*might also prompt the other students in the class to talk about their own*" (T4).

4.2.4 Simple, Polished, and Age-Appropriate

Teachers noted that their students are highly sensitive to design in a media-rich environment and are easily turned off by resources that do not align with their expectations. In general, teachers had found that their students preferred resources that looked polished, were not too mature or childish, and used neutral language. For example, T13 described their preferred resource "*as simplistic as possible and not super*

wordy... And it also needs to look polished... [Kids] are very dismissive... these are kids who are bombarded with media all the time. So, if it looks like it was done ten years ago, they're out."

Teachers reported gravitating toward games or turning a static resource into an activity to increase engagement. For example, T12 said: "*I would copy, paste this into a little checklist, like go through and check off maybe one thing you learned. You know... it's just a handout. I would turn it into an activity.*"

Teachers emphasized the importance of limiting the amount of written content in the resource and also how much writing the resource required students to do. In addition, teachers noted that it was also a deterrent to them. "*... if I'm reading a two-page document to find out what the lesson is*" said T14, *I'm not going to use it*." The tone of the resource should not come across as "*preachy*" (T15) or "*talking down to them*" (T13). Teachers are aware that students may feel judged by a resource that has an overly prescriptive tone and may become defensive and "*tune [it] out*" (T15) as a result.

4.2.5 Non-technical

Surprisingly, teachers had reservations about using resources that require the use of technology in the classroom because it creates many challenges and barriers for teachers. For example, not all schools have the ability to offer a one-to-one ratio of Chromebooks for students to use, which means that students have to share computers. This limited their ability to optimally engage with some type of resource, such as computer games.

Teachers also noted that resources designed with an overly technical focus can make them less usable to teachers. They reported rejecting or making modifications to the resources due to the lack of perceived appropriateness of the resource for their class, such as the correct literacy level whether they had the means to incorporate the resource into their classroom (e.g., number of available tablets to access the resource). Further, resources that required user account registration created significant barriers because having to remember multiple account details and logging in before each lesson is a hassle.

4.2.6 Modular and Adaptable

Teachers reported that they prefer to approach cybersecurity and privacy subjects in a flexible way. Therefore, they preferred resources that provided options to adapt the resource to accommodate their constraints and needs. These included modifying a resource, adjusting the length of the lesson, and making the lessons more accessible without technology.

Materials that included multiple smaller lessons packaged around a topic or educational outcome were preferred. Teachers noted that having multiple topics to choose from was helpful in offering them a "*starting point*" (T10). Further,

they appreciated being able to choose one or two pieces from a package of resources, rather than feeling constrained by a single resource or pressured to use a resource in its entirety.

4.2.7 Trusted and Free

In choosing resources, teachers considered the reputation of the organization making the resource. Their trust in the organization was mainly determined by the professional look of the resource and the website. They also perceived resources recommended by colleagues as more trustworthy. Most used free resources because they do not have a budget through their school to buy materials.

4.2.8 Fit Within Current Practices

Teachers reported seeking resources that they could easily incorporate into their teaching practices and responsibilities, such as how well the material met their curricular needs. Many felt that they did not have enough time to appropriately cover what they already have to teach in the curriculum. As a result, teachers are likely to dismiss resources that do not clearly outline how they connect to the existing curriculum.

Teachers gravitated to resources that clearly outlined the learning objectives and success criteria, noting that this helped them with their administrative responsibilities: "*learning goals is a big thing with our school board,* said T11, *"you have to state what your learning goal is, what your success criteria is...*"

4.3 Assessments

In general, teachers reported using a variety of measurement strategies with their students, although they had clear preferences for the type of assessment. Teachers reported overall positive outcomes from their lessons, but noted the lack of clarity around what effects of their lessons had on their students and the long-term learning outcomes. These concepts around measurement strategies and lesson results are explored further below.

4.3.1 Informal Assessments

The majority of the teachers preferred informal assessment strategies, such as relying on discussions and "*vibes*" (T8) to assess student understanding intead of using formal assessment tools such as quizzes and assignments. They opted for informal methods of assessment because they did not want to create anxiety or stress for their students due to the personal and potentially sensitive nature of cybersecurity and privacy. Assessments could also distract students from the central issue. T9 explained:

> *I don't want to grade a student on their response to something like this, because first of all, a lot of this is sort of opinion and experience-based. So, I can't really grade*

> *them on that because that's not part of the curriculum. And then if I grade them on something sort of adjacent like, for example, a written response, and I grade them on their grammar or something, then they're a lot more focused on that than the actual issue.*

Therefore, teachers felt that applying a grade did not represent the best pedagogy for teaching cybersecurity and privacy.

4.3.2 Critical Reflection

Teachers highlighted the importance of reflection in their assessment strategies. As such, they preferred assessments that facilitated critical reflection over those that measured correctness, such as multiple-choice questions. T13 explained:

> *... it's the sheer volume of media that they're consuming. It doesn't allow for reflection. It doesn't allow for you to think. It's just constant. So they don't slow down and think about it very often. And so any time that we can get them to slow down and think about what they're doing it's a win.*

Teachers also emphasized the importance of reflection for young people that extended beyond the classroom.

4.3.3 Short-Term vs. Long-Term Impact

Teachers reported mostly positive reactions to their lessons, but had mixed results when it came to seeing a lasting change in student attitudes and behaviour.

In most cases, teachers noted that their students responded positively to lessons with the immediate result being that students were eager to engage in discussions about these topics. Despite positive short-term engagement, teachers found it difficult to tell if their lessons had a lasting impact on student attitudes and behaviours. T2 noted that their presentations often ended with students self-reporting "*deleting their Facebook account*" or "*keep[ing] their eye out for activities or if their friends are acting strange.*" However, T2 and other teachers noted that this was not something they could verify.

Complicating this issue further is that it is becoming increasingly difficult for both teachers and guardians to keep track of the ever-growing number of games and online services that their student have access.

> *It is almost impossible... And you just have to hope that you've laid enough of a foundation by the time they get to that point that they're going to talk to you about it. But in most cases, they don't. And so it's a really powerless feeling... (T13)*

This highlights a unique challenge for teachers: to know when to intervene or whether their lessons are having an impact. As such, this may be a significant contributor to why most teachers reported having a reactive approach to addressing cybersecurity and privacy risks with their students.

5 Creators' Perspectives

The resource creators in our study had experience creating resources covering a wide variety of topics: 88% indicated their resources taught authentication and privacy, and 63% addressed online ethics. Fewer had created materials covering more sensitive topics such as sexting (38%), sexploitation (25%), or pornography and gambling (13%). Most of the creators in our study said that they had experience developing lesson plans (88%), learning modules (88%,) and text-only resources (88%). Some had created animated videos (63%) and web-based games (38%) . Only 25% had experience creating mobile games, comics, and gamified activities.

5.1 Curation

Resource creators shared that their first step in creating a resource is research to help them better understand what contributes to the problem and where there are gaps that their materials need to fill. However, we found that most relied on ad hoc processes to determine the topics they covered and using a variety of sources to gather evidence to support their advice due to a lack of centralized knowledge and funding bodies to support cybersecurity education. C2 noted their process for curating resources:

> *So, the topics were picked based on what the biggest issues for those were. In terms of specific aspects of fraud and things like that, we go to the statistics and we talked to the Canadian Anti-Fraud Center... We try and get an idea of what the larger problems were and then build out units around that. It's very hard to get an idea of what basic cybersecurity is because a lot of the places that provide that kind of information aren't the kind of institutions that can also provide the evidence...*

These quotes highlight how the lack of clarity around the most pressing problems and how best to address them complicates creators' processes for determining appropriate topics and creating evidence-based materials. The fact that there is no centralized place for validated information coupled with a rapidly changing technology landscape makes it harder for creators to engage in efficient processes and risks their providing outdated or outright bad advice.

Organizations, particularly not-for-profits, generally focus on "hot" cybersecurity and privacy topics to attract funding, and funding for the project limited the resources they could create. As C1 explained:

> *It is either what we can attract funding for, or alternately when we consider something to be a priority, we find time to do it. Obviously, that's more practical with something like a tip sheet or a lesson plan than something like a video or something more, that has more hard costs or money out the door. So, what we kind of do is we try to match funding opportunities with things that we want to do, and we do that in a variety of different ways.*

5.2 Processes and Methodologies

Once a project plan or funding was secured, the creators reported a mix of activities, including engaging stakeholders, developing partnerships, bringing in subject matter experts, prototyping, reviewing, and then launching and promoting their products. In several cases, creators also hired translators to convert their materials into French.

While some creators used existing theories and academic practices (*e.g.*, participatory design, Agile, and user-centred design) to inform the development of their resources, others did not follow any established design methodology or framework. For example, C7 explained why they avoided using frameworks in the development of their resources:

> *I probably couldn't name a framework for you. How about that? I was a teacher for 38 years and a curriculum designer and I know there are frameworks for doing that. But you know what we've discovered over the years? Those frameworks get in the way of being productive. And as soon as you say framework, that means, okay, there are rules, this is the way we go. And that really limits these trips to the side that generate some serious fruit. And so what we did, we just went and just everything was on the table. And then we sift through it afterwards.*

In general, creators used broad terms to describe the effectiveness of the resource, such as "engaging", "usable", and "accessible". They spoke of concerns around the explainability and transmissibility of the material, with a focus on making the content understandable to audiences beyond its initial stakeholder group. They also mentioned concerns about knowledge transfer to apply the acquired knowledge to new situations and presenting authentic learning opportunities where students engage their problem-solving skills.

5.3 Design

We found that resource creators acknowledged many of the same high-level factors as teachers when discussing how resources were chosen. Creators discussed optimizing the design of their resources to suit the expectations of the students, such as incorporating modern design aesthetics to capture their interests and engagement during lessons. Further, creators highlighted the importance of age-appropriate design and communication in the design of their materials, many of which matched teachers sentiments. These included ensuring that the materials had "*fun and engaging branding*" (C5) to appeal to students' aesthetic tastes, have minimalist writing to make sure the resource isn't too "*text-heavy*" (C6), to ensure that the resource is at the right literacy level, and provide opportunities to develop skills for *"critical thinking [and] ethical decision-making"* (C1).

Like teachers, creators also acknowledged the importance of stories, analogies, and metaphors as educational tools. To

address this need, creators reported creating resources such as articles, comics, and games with a specific narrative focus. Creators also showed an awareness of the importance of tone in their resources, several highlighting that traditional advice had focused on "*only teaching the bad*" (C3) and understood that there is a growing need to balance negatives with the positives of technology use. Furthermore, one creator noted the importance of not being prescriptive in their advice and seeing their materials as "*a basis for a conversation*" (C2) so as not to shut down the communication channels between young people and educators. Creators were aware of the importance of keeping their materials up-to-date for teachers, despite this being a significant challenge for their organizations due to limited funding and resources.

Overall, the creators highlighted many of the same concerns and considerations as teachers, and generally showed an alignment of understanding with teachers' needs and constraints in the designs of the educational resources.

5.4 Evaluation

Resource creators overwhelmingly reported that teachers and students are difficult to reach or work with due to teachers being "*overwhelmed with the amount of work*" (C4), and students being a vulnerable stakeholder group that requires additional risk management and approval processes. This led to reliance on proxies, such as someone who worked closely with teachers rather than directly working with teachers, or involving teachers only near the end of the design process. C6 shared their struggle:

> *So doing something like, you know, a user-centered design process where the teachers were on the design team was just not in the cards. And we also had decided, you know, that was not something that was needed because we did have people on our team who work very, very closely with these teachers. And so they could kind of be their advocates. And again, they were former teachers... So yeah for the majority of the design process, they were our advocates for the teachers. We were not directly talking to the teachers... So we really started involving teachers at the end when the final product was ready. So when the high-fidelity prototype was completed, that's when I did a study with teachers.*

One risk of involving stakeholders at the end of the process is that it constrains what teachers can offer feedback on and missing important problems or opportunities that needed to be addressed near the beginning of the process.

Creators also wished to improve the measurement of the effectiveness of their resources by conducting more frequent and in-depth evaluations. However, due to limited funding and constraints on their time, the majority of creators do not evaluate their resources or used informal methods, such as soliciting opinions conversationally after presenting their resources to a small group of stakeholders. Furthermore, creators re-

ported that they primarily focused on asking self-reports of behavior change in their evaluations, rather than on learning outcomes. Creators were concerned with their inability to measure whether there were long-term changes in behavior, the ecological validity of the materials they were creating.

6 Discussion

We conducted two qualitative interview studies examining how educational resources for teaching cybersecurity are being used and evaluated by teachers, and how they are being designed and distributed by creators. We interviewed 15 Canadian teachers about their experiences teaching cybersecurity and privacy in the classroom, and 8 creators about their experiences creating cybersecurity resources. We then conducted a thematic analysis of their responses.

From our analysis, we found that teachers were using predominantly lesson-oriented resources in their teaching which they generally found to be effective. Further, their considerations in deciding on resources focused on how well the resources aligned with their teaching needs and how engaging and effective they thought it would be for their students. The interviews further highlighted that teachers are predominantly approaching these topics in a reactive and ad hoc way which impacts their process for finding and choosing resources, and measurement strategies when teaching these topics to their students.

Creator interviews showed that creators had a generally good understanding of what teachers want and need from the resources they are creating. However, when investigating their processes for designing and disseminating their resources, we found inefficiencies as well as a mix of organizational and external constraints that limited their ability to engage in best practices.

6.1 Different Educational Approaches

It quickly became clear in our interviews that teachers were approaching topics in cybersecurity and privacy not from a technical perspective, but from a perspective framed around safety. This shaped what kind of resources they chose, how they approached teaching, and how they evaluated students.

Teachers often described taking a reflective approach to teaching security and privacy, and choosing teaching strategies that emphasized critical thinking and inquiry. Teachers sought to connect the material to students' lived experiences, often by approaching these topics reactively. Teachers frequently described teaching strategies such as class discussions, and emphasized the importance of candidness and students' emotional safety in these discussions. Teachers adapted existing materials to fit these reflective teaching modalities.

Rather than starting with technical strategies and integrating more personal impacts of the material from there, teachers

expressed a preference for using stories and role-play to encourage students to explore the ways in which their digital footprints might affect them. Teachers said that strategies such as scare tactics or presenting shocking information were often good ways to get students engaged in the material, and contrasted these "shock" techniques with maintaining an open and honest rapport with students that would enable honest and safe discussion. This emphasis also led to teachers adapting more technical material to work with their narrative-focused strategies.

Teachers expressed a clear preference for informal assessments for cybersecurity, privacy, and digital literacy topics. Much of this had to do with the style of teaching, and the method of approaching these topics, which did not lend themselves to formalized assignments or quizzes. The majority of teachers preferred informal evaluation strategies, and relied more on discussion, and engagement as metrics for the success. Teachers were clear that the subject matter itself was a source of stress for their students, and were reluctant to compromise or complicate the classroom tensions by adding formal assessment items.

In our interviews, creators rarely brought up these kinds of considerations about what kind of educational approach to take, or framed cybersecurity education as part of a conversation or situation outside of a dedicated lesson. While it is possible that they are aware of them, they did not seem to frame their approach to designing lessons with the same considerations. We suggest that if creators had a greater awareness of the constraints and considerations affecting teachers this could help them create resources that better served these approaches.

6.2 Conflicting Processes

Educators and resource creators approach the same problem from different perspectives: how can cybersecurity topics be best synthesized for delivery to students? However, in analyzing our interview data, we noticed that creators and teachers were approaching their task from different angles. Creators were using a top-down process, starting with trends in cybersecurity topics, funding considerations, and other high-level factors to consider the design of security resources. Teachers were more likely to be starting with bottom-up factors that reflected the realities of their teaching context, such as student safety and curriculum demands.

In our interviews, creators tended to start with more of a blank slate when considering the design and creation of resources. Creators brought up some constraints relating to factors such as funding, but in general, approached the design of resources from a perspective framed around the cybersecurity topics. Once a project plan or funding was formalized, creators reported a mix of activities, including engaging stakeholders, developing partnerships, bringing in subject matter experts, prototyping, reviewing, and then launching and pro-

moting their products. Few of these activities involved direct feedback from teachers or students.

When teachers described how they chose resources, they mentioned a variety of factors. Many of these factors were directly related to emergent events in their classroom: instances of bullying or other conflict, interpersonal relations between students, students' digital lives and presence, events happening in the local community, etc. Teachers were also driven by contextual factors such as curriculum demands, the other material they were teaching, and the time available to them. As a result, teachers were likely to pick and choose pieces of resources, using a bottom-up technique to assemble material that suited these constraints. Their teaching tended to be reactive, rather than proactive, and their resource discovery strategies were broad. Instead of beginning with the resource packages made available through creators, they tended to begin with Google searching. Teachers also described addressing cybersecurity topics in non-technical classes (*e.g.,* health class), often because they afforded the time and discussion needed to approach topics in a way that was customized to teachers' students.

Although teachers expressed few complaints about the resources they were using, it seemed clear that these resources were not particularly created with their constraints in mind. One effect of the mismatch seemed to be that teachers were forced to de-prioritize cybersecurity topics in comparison to other curriculum topics. In our interviews, teachers suggested that having cybersecurity topics explicitly tied into other courses, particularly math and languages, would allow them more opportunities to engage with the material. Because of their reactive approach to teaching cybersecurity and privacy, teachers were in search of materials that they could easily fit into both their current class plan and their curricular mandates. To facilitate this, teachers generally reported using one-off lesson materials rather than resources that required a series of lessons that required building off on previous lessons and which would take multiple classes to cover.

We suggest that a better alignment between these two processes might help the development of resources that are more effective for teachers. Possibly, creators are aware of this mismatch – in our interviews, creators expressed frustration with the lack of communication with teachers and students. However, they also acknowledged the lack of a formal design process and methodology. Using a design process that prioritized direct involvement with teachers at early stages of the projects could help this mismatch, and better influence both the format and content of resources.

6.3 What is Available and What Gets Taught

Another way in which the differences in approach between teachers and creators became apparent was in the design and format of the resources themselves. The resources and materials created and disseminated by creators do not match

teachers' classroom constraints and needs.

Our research suggests that few available cybersecurity resources are taught to students in the classroom due to their incompatible formats or lack of flexibility in adapting the material to classroom teaching. We found that teachers relied mainly on lesson plans, short live-action and animated videos, and learning modules to teach cybersecurity and privacy. However, previous research [39] has found that most of the available cybersecurity educational tools are games and videos. Learning modules represent only around 8% of all available resources, and lesson plans are usually supplementary material to support other types of multimedia learning (but are not always available) [39]. We found that resources such as tabletop games, mobile games, comics, and gamified activities were rarely used by our teacher participants. While computer games were sometimes used for teaching purposes, creators may be overestimating their usefulness in classroom environments compared to other non-interactive material like text-only resources, which we found to be used almost as often as games.

Creators and teachers are key stakeholders in determining *what* and *how* cybersecurity and privacy topics should be taught. As curators of learning resources, resource creators communicated a lack of guidance on what cybersecurity problems should be prioritized, leading to confusion about what topics to cover. We also found that the topics creators support are sometimes constrained by lack of funding opportunities to support the development of the learning resource. Therefore, the disseminated resources may not always address security and privacy issues faced by children and youth on the ground or align with the topics that teachers need to cover in the classroom. Creators referenced blind spots in the development process, such as who they are designing for and the underlying need the material is trying to address. "*We'll just use the phishing example because most of the time we get requests around how do students be more aware of malicious attacks or phishing,*" C5 declared, "*But then we don't really understand who is this going to... what are the students really experiencing? How are they digesting that information? ...*" We found that creators' resource development efforts focus on design considerations to make resources more attractive and engaging for young people, which could lead creators to develop certain types of resources over others, such as games and videos [39]. Other stakeholder perspectives are also present in the design process, such as that of school boards and funding bodies, but our interviews suggest that creators were prioritizing the learning needs of their primary audience (i.e., children and young people), but the need to support teachers was usually not considered until the end of the design process (if at all).

Compared to creators, we found that teachers prioritized suitability to their teaching needs in conjunction with the learning needs of their students. For example, teachers sought resources that could easily fit into their current teaching plan

and curricular mandates. To facilitate this, teachers generally reported using one-off lesson materials rather than scaffolding a series of lessons that could take multiple sessions to cover. Our results also suggest that teachers may deliberately avoid teaching certain topics that they consider sensitive and uncomfortable, such as sexting, or technical topics on which they lack expertise. This indicates that more careful curation of topics from creators is required to support teachers' needs. Our interviews suggest that simple, flexible, and modular resources like short videos, adaptable learning modules, and lesson plans for facilitating classroom discussions are easier to use by teachers than resources that require more complicated setups and time commitment. A closer relationship between teachers and creators in the design phase would likely help address many of these issues.

7 Conclusion

As online resources are entangled more and earlier into childrens' lives, the importance of effective education in cybersecurity and privacy continues to grow, bringing with it the need for well-designed and effective resources for teaching these topics. In this work, we explored how existing resources align with the needs of teachers using them. We conducted two qualitative interview studies with 15 teachers and 8 resource creators. We found that teachers approached cybersecurity and privacy from a safety-oriented rather than a technical perspective and often did so as an ad hoc reaction to external events in the classroom, school or community. As a result, they preferred informal assessment strategies like facilitated discussions over formal assessment methods like tests. Resource creators generally had a good understanding of the learning needs and interests of their students, but generally did not prioritize their design considerations of the resources for teachers' delivery of the material in the classroom. As a result, our findings suggest that teachers access and use only a small portion of the cybersecurity educational content available to instruct children due to their rigid and incompatible formats to adapt the material for classroom teaching. Specifically, computer and mobile games—the most widely available type of cybersecurity educational resource—are rarely used in classroom teaching contexts. In contrast, teachers are more likely to use modular lessons that can be easily adjusted to their teaching using resources such as lesson plans, short instructional videos, and segmented learning modules. We suggest that better integration of the factors affecting teachers into the resource creator processes could enable more flexible lessons that are more widely applied in the classroom, resulting in better knowledge of cybersecurity and privacy for students.

8 Acknowledgments

References

[1] Media Education in Ontario. https://mediasmarts.ca/teacher-resources/digital-media-literacy-outcomes-province-territory/media-education-ontario, Jan 2012. 2022-04-15.

[2] Canadian Cybersecurity Awareness Stakeholders Teleconference Report. https://www.serene-risc.ca/public/media/files/prod/page_files/27/SETA-Conference-Report-FINAL.pdf, January 2020. 2022-03-12.

[3] Cybersec 101. Cybersec101. https://www.cybersec101.ca, 2016. 2022-06-22.

[4] Osman Sirajeldeen Ahmed, Saeed Ameen Nasef, Alaa Zuhir Al Rawashdeh, and Mohd. Elmagzoub Eltahir. Teacher's awareness to develop student cyber security: A Case Study. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(10):5148–5156, 2021.

[5] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006.

[6] Kara Brisson-Boivin. The Digital Well-Being of Canadian Families. Media Smarts, 2018.

[7] David Byrne. A worked example of Braun and Clarke's approach to reflexive thematic analysis. *Quality & Quantity*, 56(3):1391–1412, 2022.

[8] Wen-Yen Chiu and Hsuan-Fu Ho. Time to Educate the Educators: An Evaluation of Cyber Security Knowledge Awareness and Implementation for School Teachers in Taiwan. In *Proceedings of International Conference on Technology and Social Science (ICTSS 2019)*. Atlantis Press, 2019.

[9] Common Sense Education. Compliance training: Protecting student privacy for teachers, 2023.

[10] Isabella Corradini and Enrico Nardelli. Developing Digital Awareness at School: a Fundamental Step for Cybersecurity Education. In *International Conference on Applied Human Factors and Ergonomics*, pages 102–110. Springer, 2020.

[11] Katie Davis and Carrie James. Tweens' conceptions of privacy online: Implications for educators. *Learning, Media and Technology*, 38(1):4–25, 2013.

[12] Facebook. Youth Portal. https://www.facebook.com/safety/youth. 2022-07-17.

[13] Google for Education. Teach and Learn Practical Digital Skills - Applied Digital Skills. https://applieddigitalskills.withgoogle.com, 2022. 2022-07-17.

[14] KnowledgeFlow Cybersafety Foundation. Curriculum Creation | KnowledgeFlow Cybersafety Foundation. https://knowledgeflow.org/solution/curriculum-creation. 2022-07-18.

[15] Google. Be Internet Awesome. https://beinternetawesome.withgoogle.com. 2022-07-17.

[16] Tea Hadziristic. *The State of Digital Literacy in Canada: A Literature Review*. Brookfield Institute for Innovation Entrepreneurship Toronto, Canada, 2017.

[17] Michael Hoechsmann and Helen DeWaard. USE, UNDERSTAND & CREATE: A Digital Literacy Framework for Canadian Schools. 2022-03-12, 2015.

[18] Annalise Huynh and Nisa Malli. *Levelling up: The quest for digital literacy*. Brookfield Institute for Innovation Entrepreneurship, June 2018.

[19] iKeepSafe. Data Privacy in Education: An iKeepSafe educator training course, 2016.

[20] Carrie James, Emily Weinstein, and Mendoza Kelly. Teaching Digital Citizens in Today's World: Research and insights behind the Common Sense K–12 Digital Citizenship Curriculum. 2022-07-14, 2021.

[21] Giti Javidi, Ehsan Sheybani, and Zacharias Pieri. A Holistic Approach to K12 Cybersecurity Education. In *Proceedings of the International Conference on Frontiers in Education: Computer Science and Computer Engineering (FECS)*, pages 77–80. ProQuest, 2019.

[22] Priya C. Kumar, Marshini Chetty, Tamara L. Clegg, and Jessica Vitak. Privacy and Security Considerations for Digital Technology Use in Elementary Schools. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13. ACM New York, NY, USA, May 2019.

[23] Sana Maqsood. *The Design, Development, and Evaluation of a Digital Literacy Game for Preteens*. PhD thesis, Carleton University, Ottawa, January 2020.

[24] Sana Maqsood and Sonia Chiasson. "They think it's totally fine to talk to somebody on the internet they don't know": Teachers' perceptions and mitigation strategies of tweens' online risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–17. ACM New York, NY, USA, 2021.

[25] Florence Martin, Tuba Gezer, and Chuang Wang. Educators Perceptions of Student Digital Citizenship Practices. *Computers in the Schools*, 36(4):238–254, 2019.

[26] Common Sense Media. Common Sense Media: Age-Based Media Reviews for Families | Common Sense Media. https://www.commonsensemedia.org. 2022-07-17.

[27] Microsoft. Digital literacy courses, programs and resources | Microsoft Digital Literacy. https://www.microsoft.com/en-us/digital-literacy. 2022-07-17.

[28] James Nicholson, Julia Terry, Helen Beckett, and Pardeep Kumar. Understanding Young People's Experiences of Cybersecurity. In *European Symposium on Usable Security 2021*, pages 200–210, 2021.

[29] Ontario College of Teachers. Discipline Decisions | Ontario College of Teachers. https://www.oct.ca/public/complaints-and-discipline/decisions. 2022-10-19.

[30] Teaching Privacy. Teaching Privacy. https://teachingprivacy.org. 2022-06-07.

[31] Portia Pusey and William A. Sadera. Cyberethics, Cybersafety, and Cybersecurity: Preservice Teacher Knowledge, Preparedness, and the Need for Teacher Education to Make a Difference. *Journal of Digital Learning in Teacher Education*, 28(2):82–85, 2011.

[32] Farzana Quayyum, Daniela S. Cruzes, and Letizia Jaccheri. Cybersecurity awareness for children: A systematic literature review. *International Journal of Child-Computer Interaction*, 30:100343, 2021.

[33] Nurul Amirah Abdul Rahman, Izzah Hanis Sairi, Nurul Akma M. Zizi, and Fariza Khalid. The Importance of Cybersecurity Education in School. *International Journal of Information and Education Technology*, 10(5):378–382, 2020.

[34] Teaching Security. Teaching Security. https://teachingsecurity.org. 2022-07-18.

[35] Media Smarts. Media Smarts | Teacher Resources. https://mediasmarts.ca/teacher-resources, November 2011. 2022-06-20.

[36] Student Privacy Compass. Student privacy training for educators, 2023.

[37] Leah Zhang-Kennedy, Khadija Baig, and Sonia Chiasson. Engaging Children about Online Privacy through Storytelling in an Interactive Comic. In *Electronic Visualisation and the Arts (EVA 2017*, pages 1–11, July 2017.

[38] Leah Zhang-Kennedy, Robert Biddle, and Sonia Chiasson. Secure Comics: An Interactive Comic Series for Improving Cyber Security and Privacy. In *Proceedings of the 31st British Computer Society Human Computer Interaction Conference*, Swindon, GBR, 2017. BCS Learning & Development Ltd.

[39] Leah Zhang-Kennedy and Sonia Chiasson. A Systematic Review of Multimedia Tools for Cybersecurity Awareness and Education. *ACM Computing Surveys (CSUR)*, 54(1):1–39, 2021.

[40] Leah Zhang-Kennedy, Sonia Chiasson, and Robert Biddle. The Role of Instructional Design in Persuasion: A Comics Approach for Improving Cybersecurity. *International Journal of Human-Computer Interaction*, 32(3):215–257, February 2016.

9 Appendix A: Teacher Pre-Interview Questionnaire

1. What is your gender? *[Multiple choice] (Male, Female, Self-identify [textbox])*

2. What is your age bracket? *[Multiple choice] 20–29, 30–39, 40–49, 50–59, 60+*

3. What did you study in university? *[Textbox]*

4. How long have you been teaching? (in years) *[Textbox]*

5. What subjects have you taught? *[Textbox]*

6. What grade(s) do you teach? *[Textbox]*

7. Where do you teach? (e.g., School name and district) *[Textbox]*

8. What technologies do you use in the classroom? *[Textbox]*

9. Have you ever helped a student deal with a digital literacy or cybersecurity issue? (e.g., cyberbullying, accidentally sharing personal information) *[Multiple choice] Yes, No*

   (a) If yes, please describe the incident (without identifying the student) *[Textbox]*

10. Have you used any educational resources designed to help you teach principles of cybersecurity and privacy in your classroom? *[Multiple choice] Yes, No*

    (a) If yes, please describe what resource you have used and who created it (e.g., MediaSmarts, Teachers-PayTeachers, a colleague, etc.) *[Textbox]*

11. How would you rate your comfort with teaching the following cybersecurity and privacy factors to your students? *[Likert scales: 5 = very knowledgeable, 1 = not at all knowledgeable] General cybersecurity (spoofing, malware, pharming, passwords), E-safety, E-privacy, Digital citizenship and literacy, Data security, Phishing, Network security, Software security*

12. Do you have any comments about the previous question you would like to share? *[Textbox]*

13. What are some of the types of educational resources you used in the past for teaching cybersecurity? *[Multiple answer-multiple choice] Games (web-based or computer games), Games (Apps on mobile devices), Games (Non-digital board or tabletop games), Videos-Films, Videos-Animations, Learning modules, Comics, Text-only resources, Gamified activities (e.g., role-playing), Lesson Plans, Other (Please specify)*

    (a) For each checked resource, please list the name of the sources and include links to the resources if possible. *[Textbox]*

14. What areas are you knowledgeable about in cybersecurity and privacy? *[Likert scales: 5 = very knowledgeable, 1 = not at all knowledgeable] Authentication, Cyberbullying, Cybersecurity (software threats, spam, scams, fraud, identity theft), Extensive Internet Use, Gambling, Online Hate, Online Ethics, Online Marketing, Privacy, Pornography, Sexual Exploitation, Sexting, Other*

15. What digital issues have you taught? *[Multiple answer multiple choice] Authentication, Cyberbullying, Cybersecurity (software threats, spam, scams, fraud, identity theft) Extensive Internet Use, Gambling, Online Hate, Online Ethics, Online Marketing, Privacy, Pornography, Sexual Exploitation, Sexting, Other*

10 Appendix B: Creator Pre-Interview Questionnaire

1. What is your gender? *[Multiple choice](Male, Female, Self-identify [textbox])*

2. What is your age bracket? *[Multiple choice] (20–29, 30–39, 40–49, 50–59, 60+)*

3. What organization do you work for?*[Textbox]*

4. Please indicate the type of organization you work for: *[Multiple choice] Private sector (e.g., business), Public sector (e.g., government, academic institutions), Not-for-profit, Other (please specify)*

5. How many employees are at your organization? *[Multiple choice] 1–9, 10–49, 50–99, 100–499, 500–999, 1,000–4,999, 5,000–9,999, 10,000+, Don't know*

6. What is your most recent job title? *[Textbox]*

7. What is your highest level of education? If you are currently in school, please choose the degree that you are enrolled in. *[Multiple choice] Less than a high school degree, High school degree or equivalent, College degree, Bachelor's degree, Master's degree, Doctoral degree, Other professional degree*

8. What did you study in university? *[Textbox]*

9. Does what you study in university relate to your work as a creator of these resources?

    (a) If yes, how so? *[Textbox]*
    (b) If no, where did you learn the skills related to your work? *[Textbox] (For example, cybersecurity, privacy, and instructional design)*

10. How long have you been creating these sorts of resources? (Professionally or otherwise) *[Textbox]*

11. What are some of the types of educational resources you helped to create in the past for teaching cybersecurity? *[Multiple choice multiple answer] Games (web-based or computer games), Games (Apps on mobile devices), Games (Non-digital board or tabletop games), Videos-Films, Videos-Animations, Learning modules, Comics, Text-only resources, Gamified activities (e.g., role-playing), Lesson Plans, Other (Please specify)*

12. For each checked resource, please list the name of the sources and include links to the resources if possible. *[Textbox]*

13. What digital issues do the educational resources you helped to create address? *[Multiple choice multiple answer] Authentication, Cyberbullying, Cybersecurity (software threats, spam, scams, fraud, identity theft), Extensive Internet Use, Gambling, Online Hate, Online Ethics, Online Marketing, Privacy, Pornography, Sexual Exploitation, Sexting, Other*

14. What areas are you knowledgeable about in cybersecurity and privacy? *[Likert scales: 5 = very knowledgeable, 1 = not at all knowledgeable] Authentication, Cyberbullying, Cybersecurity (software threats, spam, scams, fraud, identity theft), Extensive Internet Use, Gambling, Online Hate, Online Ethics, Online Marketing, Privacy, Pornography, Sexual Exploitation, Sexting, Other*

11 Appendix C: Teacher Interview Guide

Teaching Practices

1. How long have you been teaching cybersecurity and privacy topics to your students?

2. Have you done any professional development in cybersecurity or privacy through your school, and if so, can you describe what was involved?

3. Have you done any professional development in cybersecurity or privacy through your school, and if so, can you describe what was involved? If not, why not?

4. Please describe your most recent experience teaching cybersecurity or privacy to your students.

   (a) Why did you decide to teach this particular lesson? (What precipitated the need to cover this topic?)

   (b) What grade were these students when you taught this material?

   (c) How did they react to the lesson?

   (d) Did you see changes in the behaviour or attitudes of your students after the lesson?

5. What strategies do you use to engage your students with these topics in the classroom?

6. How else is privacy and security being addressed in your school?

Resource Selection, Effectiveness, and Experience

1. What I would like you to do now is walk me through how you would go about finding and choosing a lesson or resource for teaching cybersecurity and privacy to your students.

   (a) What are your considerations for choosing a resource?

   (b) Where do you start your search?

2. Now I would like you to show me the 1 or 2 resources you have been using in teaching cybersecurity and privacy to your students and then I'd like to ask you a few questions about them.

   (a) How did you first learn about "X" resource? (web page, lesson plan, etc.)

   (b) What concept(s) does this resource teach?

   (c) How long have you been using "X" resources to teach this concept?

   (d) Why did you choose this particular resource to teach this concept?

   (e) What is it about this resource that you like?

   (f) Do your students seem to be engaged when you use this resource?

      i. If yes, what do they seem to like about it?

      ii. If no, what seems to impede their engagement?

   (g) What part of the design do you think make the resource particularly effective for learning about cybersecurity or privacy?

   (h) Does this resource have a teacher's facilitation guide or any support material to help explain to you how to teach it? If yes: Do you use it?

   (i) How do you incorporate the resource in your teaching? For example, have you made any modifications to the resource to make it work better for you?

      i. If you made changes, what changes did you make?

   (j) Is there an assessment component to this resource?

      i. If yes, do you use the assessment?

      ii. If not, how do you measure the effectiveness of the resource?

   (k) Are there things about this resource that you dislike or feel could be improved? If yes, how so?

3. Do you have any other feedback you would like to share?

12 Appendix D: Creator Interview Guide

Background Questions

1. Can you describe the type of work you do relating to cybersecurity education?

2. Can you describe the types of resources you helped to create and the target audience?

Process for Resource Development and Dissemination

1. Can you describe for me what types of educational resources you create?

2. Do these include supporting materials like teaching guides and assessments?

3. How do you decide what topics to base your materials on (what topics should tweens need to know)?

4. Please walk me through your organization's design process for developing cybersecurity and privacy-related educational resources.

5. Can you describe the design methodologies and/or frameworks that your organization uses for developing educational resources? (e.g., user-centered design, agile, participatory design)

6. What are the types of stakeholders you engage within the design process (e.g., privacy experts, end-users, teachers, interaction designers, developers, content writers)?

7. When and how do you engage your stakeholders during the design process?

8. Do you measure the success of your resources? (e.g., the popularity of your resources via analytics, usability testing)?

    (a) If yes, what types of data do you collect?

    (b) If yes, is there anything from the data that surprised you?

9. Do you evaluate your educational resources with teachers and students?

    (a) If yes, please describe your process and methodology for doing the evaluation.

    (b) If yes, broadly speaking, what have your results been of your tests?

    (c) Where do you feel there are opportunities for improvement in your evaluation processes?

10. Have you gotten unsolicited feedback from educators after they've used one of your resources?

    (a) If yes, what sorts of things did educators highlight in their feedback?

11. What are things you wish you knew when designing these materials?

12. What is the process for disseminating or deploying these educational resources to teachers, administrators, and students when they are done?

13. Do you have a formal communications plan?

14. How do teachers, school administrators, and students find your educational resources (e.g., browsing, direct search, recommendations, curriculum)?

15. Are there specific support or resources for helping teachers adapt the educational resources for classroom use (e.g., teaching guide)?

Improvements and Recommendations

1. What do you like about your process for creating these resources?

2. Where do you feel there is room for improvement?

3. What is one thing you would like to find out from my interviews with teachers?

4. What are your design recommendations for designing security and privacy educational tools for tweens?

5. What are your design recommendations for creating support materials for teachers to facilitate the use of cybersecurity educational tools in the classroom (E.g., teacher's guide, activity guide)?

Negative Effects of Social Triggers on User Security and Privacy Behaviors

Lachlan Moore
Waseda University / NICT

Tatsuya Mori
Waseda University / NICT / RIKEN AIP

Ayako A. Hasegawa
NICT

Abstract

People make decisions while being influenced by those around them. Previous studies have shown that users often adopt security practices on the basis of advice from others and have proposed collaborative and community-based approaches to enhance user security behaviors. In this paper, we focused on the *negative* effects of social triggers and investigated whether risky user behaviors are socially triggered. We conducted an online survey to understand the triggers for risky user behaviors and the practices of sharing the behaviors. We found that a non-negligible percentage of participants experienced social triggers before engaging in risky behaviors. We also show that socially triggered risky behaviors are more likely to be socially shared, i.e., there are negative chains of risky behaviors. Our findings suggest that more efforts are needed to reduce negative social effects, and we propose specific approaches to accomplish this.

1 Introduction

Human beings are intrinsically social. In the usable privacy and security field, researchers have found plenty of evidence that users are socially influenced when they make security and privacy (S&P) decisions [11, 12, 32, 40, 45, 48, 58]. For example, non-expert users learn lessons from S&P advice and stories from others such as family, friends, and colleagues [11, 45, 48, 49]. In such small social groups, people sometimes both receive and give S&P tech care to each other [31]. Furthermore, users can be influenced not only by people they are close with but also by strangers online.

Users sometimes ask strangers for S&P advice on forums and question-and-answer sites [23, 41].

While researchers have focused on and attempted to take advantage of the positive aspects of social effects, we should not turn away from the *negative* aspects. Negative social effects include the possibility that non-expert users may be encouraged by others to engage in risky or insecure behaviors. In the context of teenagers' health, having friends who smoke or drink, and invitations to partake in these activities from friends are the dominant factors to smoking and drinking in teenagers [34]. Does the same kind of negative chain occur in the context of digital S&P risks? Not enough systematic research has been done on the negative aspects of social effects in the S&P decision-making of non-expert users.

A popular model in behavioral psychology suggests that human behavior is a product of motivation, ability, and trigger, and *trigger* is defined as something that prompts action [21]. In 2019, Das et al. showed that social triggers were more common than proactive and forced triggers when it came to users' S&P behaviors [11]. They also showed the potential of positive social chains, where socially triggered S&P behaviors are more likely to be shared with others. In this paper, we expand their work to understand social triggers for *risky* behaviors. We examine whether researchers need to work on reducing the negative aspects of social triggers in addition to activating the positive aspects. Specifically, we address the following research questions:

RQ1 How frequent are the social triggers for *risky* user behaviors?

RQ2 By whom are users triggered to engage in *risky* behaviors?

RQ3 What are the factors of the social triggers for *risky* user behaviors?

RQ4 How often and why do users share their *risky* behaviors with others?

To address these research questions, we conducted an online survey ($N = 417$) in which we asked participants about the practices and contexts of risky behaviors. Specifically,

we asked participants to select risky behaviors that they had engaged in over the past 6 months, and we then asked them about the behavioral triggers that preceded their behavior, the person associated with the triggers, and whether they shared their behavior with others. The risky behaviors we asked about were related to passwords, account and update management, internet connections, content downloads, and social media posts. We then analyzed the frequency of the triggers for risky behaviors and practices of sharing the behaviors, as well as whether they vary by individual demographics and type of risky behavior.

We found that participants sometimes engaged in the risky behavior due to social triggers; approximately 20–50% of participants observed others engaging in risky behavior or were advised to do so before engaging in the behavior. For example, of the participants who reported having downloaded illegal or unofficial software/applications and media, 48.7% had observed others doing it. Participants observed risky behaviors not only of friends, family, and colleagues but also of online strangers. We also found that the type of risky behavior significantly affected the likelihood of social triggers (observation of others and/or advice from others). Specifically, account sharing and illegal downloading were more likely to be caused by social triggers than other risky behaviors. Importantly, we showed that participants were more likely to share their risky behavior with others when their behavior had been socially triggered. This means that there are negative chains of risky user behaviors. Users share their risky behaviors primarily because they want others to get a benefit. On the basis of our findings, we propose specific approaches to reducing the negative effects of social triggers. Our recommendations include the interventions for posts on online platforms regarding risky behaviors and security education with emphasis on risky behaviors susceptible to negative social chains and the risks.

This study makes the following contributions.

- In contrast to previous studies, we focused on negative social effects on user security and privacy. We show that users are socially triggered to engage in risky behavior. Our results suggest that more efforts are needed to reduce these effects.
- We identified the factors and sources of social triggers for risky user behaviors and the factors and reasons behind the practices of sharing them. This allowed us to discover clues to reducing the negative social effects and propose specific approaches to reducing them.

2 Related Work

We first review studies that investigated risky or insecure behaviors of non-expert users. Then, we go over studies that focused on the social effects on user S&P behaviors.

2.1 Risky User Behaviors

Contrary to security researchers' and experts' expectations, non-expert users sometimes fail to implement adequate security measures or take risky actions. Ion et al. [26] and Busse et al. [9] found that security practices that experts followed and recommended were not employed by non-expert users. Specifically, many non-expert users did not use a password manager, keep their system up-to-date, or use two-factor authentication. In terms of online data privacy, although concern about data collection and misuse is growing in general [27], most users do not read privacy policies [43], and almost half of internet users share their information publicly [30].

Researchers have studied the reasons why non-expert users engage in risky behaviors, fail to implement security measures, or fail to follow security advice. For example, Milne et al. [42] demonstrated that male, younger users, and users with low self-efficacy were more likely to adopt risky behaviors. Zou et al. found that people who are female, have relatively lower levels of education, and lack prior negative experiences and technical background were less likely to adopt security practices [61]. Additionally, Fagan et al. demonstrated that users who disregard security advice perceived the benefits of compliance and the risks of non-compliance to be lower than those who adhere to the advice [19]. Users abandoned security practices when they were perceived as low-value, inconvenient, or when users overrode them through subjective judgment [61]. Moreover, users have misconceptions about S&P technologies [3, 54, 56], and Abu-Salma et al. suggested that specific misconceptions limit user motivation to adopt secure tools [3].

2.2 Social Effects on User Behaviors

Positive aspects. Studies on sources of security advice have showed that non-expert users take security advice informally from family, friends, and colleagues, as well as from formal sources such as technical support [11, 45, 48, 49]. Rader et al. [48] and Pfeffer et al. [45] found that most users have learned lessons from stories about security incidents informally from family and friends and that these stories impact the way users think about security and their subsequent behavior. Other than people that they are close with, users sometimes ask strangers for S&P advice on forums and question-and-answer sites [23, 41].

In 2019, Das et al. [11] systematically typified the triggers that lead to S&P behavior changes. They revealed that "social triggers," where users interacted with or observed others, were most common, followed by proactive triggers, where users acted absent of an external stimulus, and last by forced triggers, where users were forced to act. They also found that participants were four times more likely to share their own S&P behaviors with others when their behaviors were socially triggered. This result suggests the possibility of a

positive feedback loop.

Kropczynski et al. [31] studied the phenomenon of "Tech Caregiving" among small social groups comprised of friends, family members, and/or colleagues. They found that tech caregiving was a fluid role, where some users both gave and received tech care, and older adults and emerging adults tended to be caregivees rather than caregivers.

Note that a digital divide of security advice exists. Specifically, Redmiles et al. found that while higher skilled users, who tend to be socioeconomically advantaged, were significantly more likely to take advice from their workplace, those who were less skilled tended to take advice from family and friends [49].

On the basis of the above interactions among users, some researchers have proposed collaborative and community-based approaches to enhance user S&P behaviors [12, 18, 32, 36, 40, 58]. For example, Das et al. confirmed the effectiveness of social-proof based interventions that encourage users to incline to explore security features by showing them that their friends use security features [12]. Krsek et al. [32] demonstrated that participants who were shown suggested S&P settings from experts and the public were significantly more likely to adhere to those suggested settings than those who saw the default Facebook settings. They did not observe a significant difference in the effectiveness of social suggestions from experts and the public. Wash and Cooper [58] conducted a field experiment involving phishing training that incorporated social stories. They demonstrated that traditional facts-and-advice are more effective when provided by security experts, but stories are more effective when told by people perceived as "like me."

Negative aspects. While usable privacy and security researchers have focused on the positive aspects of social effects, relatively few studies have focused on the negative aspects. Several researchers have discussed the potential of these aspects [13, 60]. For example, Das et al. suggested that social proof may have a negative effect on the adoption of security features for users with only a few friends who adopt the features [13]. Recently, Rader [47] featured a norm-based phenomenon called pluralistic ignorance where people engage in a behavior that they privately do not believe in or approve of because they believe that everyone else approves of it. In addition, Rader showed that social expectations influence user choices to use potentially privacy-invasive technologies. This suggests that sharing information about others' behavior is likely to backfire in a pluralistic ignorance situation.

Other researchers have analyzed risky and insecure advice on social media [4, 8, 59]. For example, Akgul et al. analyzed VPN ad videos on YouTube and found that these videos include vague and potentially misleading statements about the capabilities of VPNs and internet threats [4]. Despite the prevalence of risky and insecure advice on social media, the

extent to which users who see it adopt it has not been sufficiently investigated.

In this study, we also focus on the *negative* aspects of social effects. In particular, by incorporating our concerns with these aspects into the methodology of the study by Das et al. [11] that systematically investigated the social triggers for S&P behaviors (i.e., the positive aspects), we systematically investigate the social triggers for *risky* user behaviors (i.e., the negative effects).

3 Methodology

We conducted an online survey to quantitatively and systematically investigate the impact of social triggers for *risky* behaviors. We explain the survey design, recruitment, participants, ethics, and limitations.

3.1 Survey Design

We arranged Das et al.'s questionnaire [11] that investigated triggers for user S&P behaviors to understand triggers for *risky* user behaviors. Our questionnaire consisted of six parts: instruction, risky behavior practices, behavioral triggers, sharing practices, risky behaviors of others, and demographics. The full questionnaire is provided in Appendix A.

Part–1: Instruction. At the beginning, we explained to participants the study purpose, the compensation amount and expected time for completion, and how their data would be handled. Only those who agreed to participate proceeded to the survey. Since our study focused on risky user behaviors, we needed to reduce social desirability bias. We followed the approach of previous work [50] that investigated user lies for protecting their privacy (called "privacy lies"), which can be expected to be influenced by social desirability bias as well as risky user behaviors. Specifically, we told participants that we did not consider engaging in risky behaviors to be bad or uncommon and that we were interested in them as researchers. We then asked participants to answer honestly and accurately.

Part–2: Risky Behavior Practices. First, we asked participants which of the following six behaviors they did in the past 6 months (if any):

- Connecting to an unknown, potentially unsecured public Wi-Fi and then engaging in sensitive data exchanges, such as transmitting credit card or password details through this connection,

- Reusing the same or similar passwords for different accounts,

- Downloading illegal or unofficial software/applications and media (e.g., videos, music, and games),

- Ignoring or delaying software/application updates,

- Sharing sensitive personal information online (e.g., location-based information, real-time activities, and pictures of yourself/others) to strangers on social media,

- Sharing accounts with family, friends or others.

We selected these risky behaviors on the basis of the previous work we reviewed in Section 2.1 and our discussions. Specifically, we selected risky behaviors that could occur on a daily basis and that could apply to all users, regardless of the device they own or the service they use. While these behaviors *potentially* expose users to S&P harms and are considered representative of risky behaviors that are expected to be prevalent among users, it is important to note that these risky behaviors do *not always* pose an S&P threat to users. The riskiness of each behavior is described below.

- Connecting to public Wi-Fi poses significant risks due to the potential for sensitive personal information to be collected and leaked [5]. Unsecured networks can be exploited by attackers through man-in-the-middle attacks or malware distribution. However, these risks can be mitigated by using a VPN or accessing the network through a virtual machine.

- Reusing passwords across multiple accounts increases vulnerability to cross-site password guessing attacks [10], potentially granting attackers access to sensitive information. However, this risk is minimized when password reuse is limited to inconsequential "throwaway" accounts with no sensitive data.

- Downloading illegal or unofficial software, applications, and media often introduces malware, viruses, or spyware that can compromise device security and functionality, and it may also result in legal penalties. However, these risks can be mitigated by using virtual machines or sandboxes and by downloading from reputable open-source communities or platforms.

- Neglecting or delaying software updates enables attackers to exploit known vulnerabilities [37]. While not all updates enhance security (e.g., UI updates), many do address newly discovered vulnerabilities. Additionally, vendors sometimes provide insufficient explanations in their release notes (e.g., fixing a vulnerability without explicitly stating it) [15]. Therefore, delayed updates can result in security risks, such as information leakage.

- Sharing personal information online can lead to harassment, stalking, identity theft, and physical crimes if it reveals that the user is not home [28, 46]. However, these risks are reduced when information is shared within trusted groups and privacy settings are properly configured on social media.

- Sharing accounts with family, friends, or others increases the risk of compromised security due to poor practices by other users [39]. However, some platforms mitigate

this risk by offering features such as one-time login passwords, eliminating the need to share permanent credentials.

Part–3: Behavioral triggers. For each risky behavior that participants reported engaging in, we next asked them to select the event that preceded their behavior (if any). The options were "I observed/heard about other people doing this," "Other people advised to do this," "My organization required me to do this," "Other (please specify)," and "Nothing in particular happened." Although participants could select more than one option, we considered only "Nothing in particular happened" to be an exclusive option (i.e., they could not select this option and other options at the same time).

We selected the above 5 options that can be applied to triggers for risky behaviors from the 13 options of Das et al.'s study [11] (i.e., triggers for S&P behaviors). We categorized the triggers into three higher level categories of triggers: social ("I observed/heard about other people ..." and "Other people advised ..."), organizational ("My organization required me ..."), and voluntary ("Nothing in particular happened"). Some participants selected "other" and provided text, all of which was related to voluntary decisions, such as decision-making for convenience, and was not related to social and organizational triggers. Therefore, we counted these as voluntary triggers.

If participants selected social triggers, we asked the participants about their relationship to the person whose risky behavior they had observed/heard about or who had advised them. The options included friend, family, colleague, online stranger, and media. If participants received advice from others, we also asked them if the person told them about the risks of the behavior.

Part–4: Sharing practices. For each reported risky behavior, we asked participants whether they shared their behavior with others. If they did share, we asked them to specify with whom (friend, family, colleague, online discussion, and/or other) and why. The options for the reasons for sharing include "I wanted them to get the benefits" and "I wanted them to know that I have knowledge." Participants could select multiple relationships and reasons. We also asked participants who did not share their behaviors why.

Part–5: Risky behaviors of others. We also asked participants about what percentage of the public they thought engaged in each behavior. Participants could specify a number from 0 to 100 using a slider bar.

Part–6: Security attitudes and demographics. While Das et al. [11] modeled user S&P behaviors using SeBIS (the security behavioral intention scale) [16], we adopted SA-6 (the security attitude scale) [20], which was proposed after SeBIS. We believe that attitudinal indicators are more appropriate than behavioral intention indicators for modeling

risky user behaviors. We then asked a series of demographic questions redarding their age, identified gender, education, IT knowledge, and country of residence. We included a simple attention check (a check that does not contain a trap question but simply specifies the option that participants must select) in the middle of the questionnaire.

At the end of the survey, we asked participants if they answered honestly, following the previous studies [7, 35]. We told participants that they would not be penalized/rejected if they indicated dishonesty.

3.2 Recruitment and Participants

We recruited participants through Prolific in January 2024. We advertised our survey as "a study on online behaviors" without using S&P-related terms to avoid self-selection bias related to S&P on the task-list screen. Participants were required to reside in the U.S. and be 18 or older. We used Prolific's representative-sample tool to increase the diversity of our participants. Prolific's representative sample provides a balanced sample in terms of gender, age, and ethnicity based on U.S. Census data. Prior to main data collection, we conducted pilot surveys with 31 Prolific workers to evaluate our survey design and estimate the time required for completion.

We excluded 29 participants who failed the attention check, completed the survey in less than 90 seconds, selected "no" to the honesty-check question, and/or provided incoherent responses. We finally obtained a total of 417 valid survey responses. Participants who completed the survey were compensated with $1.75, and the median completion time was 314 seconds ($20.1/hour; this amount is sufficiently higher than the U.S. minimum wage).

Table 1 shows the demographics of our participants (N = 417). Our participants were 18 to 83 years old (mean 45.5, SD 15.6), 51.3% of them identified as female, and 1.9% selected "non-binary/third gender" or "prefer not to say." In terms of knowledge in IT or related fields, 55.9% rated themselves as "strongly agree" or "somewhat agree" and 19.4% as "neither agree nor disagree." Figure 1 indicates the distribution of the SA-6 score of our participants. The mean score was 20.2 (SD 5.1).

3.3 Ethical Considerations

We carefully designed our survey design, and it was approved by the Institutional Review Board (IRB). Except for the Prolific IDs, which were necessary for compensating the participants, we did not collect any personally identifiable information. We handled all data confidentially. Participants could drop out at any time. All participants who completed the survey were compensated regardless of the quality of their response.

Table 1: Participant demographics (N = 417).

		N	%
Age	18–29	92	22.1%
	30–39	69	16.5%
	40–49	71	17.0%
	50–59	81	19.4%
	60–69	86	20.6%
	70+	18	4.3%
Gender	Male	195	46.8%
	Female	214	51.3%
	Other / Prefer not to say	8	1.9%
Education	High school	131	31.4%
	College	51	12.2%
	Undergraduate	150	36.0%
	Post-graduate	76	18.2%
	Other / Prefer not to say	9	2.2%
IT knowledge	Yes*	233	55.9%
	No	184	44.1%

*For simplicity, we show the percentage of participants who selected "strongly agree" or "somewhat agree" on a 5-point Likert scale in this table.
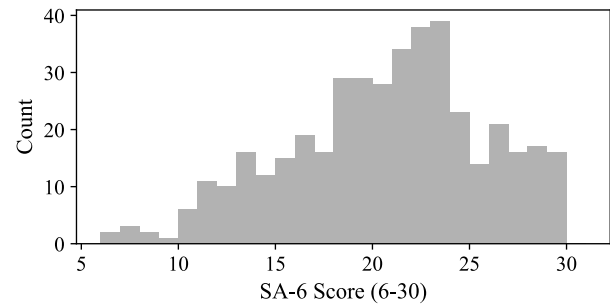
Figure 1: Distribution of SA-6 score of our participants.

3.4 Limitations

Measurement of behavioral triggers. Our study has several limitations in common with Das et al.'s study [11] in investigating behavioral triggers. In the same way as their study, we asked participants what happened before they engaged in the behavior, rather than what triggers influenced their behavior. We focused on the triggers that users perceived in the moment. Other long-term triggers, such as social norms and cultural attitudes, may influence users. Additionally, although multiple triggers may affect users, we did not ask participants which one affected them the most in consideration of recall bias. Therefore, it was not possible to quantify the strength of the impact of each trigger. Furthermore, we analyzed participants' self-reported behaviors and triggers.

Measurement of triggers for *risky* behaviors. Our study also has unique limitations in terms of measuring the behavioral triggers for *risky* behaviors. First, risky behaviors are

considered to be heavily influenced by personal motivations (e.g., the desire to take the easy way out and the desire to watch illegal content), and it is not easy to encompass the behavioral triggers that lead to risky behaviors. In fact, we initially conducted a pilot survey to typify behavioral triggers for risky behaviors, but most of the responses were about such personal motivations. Therefore, as a first step, we focused on understanding the prevalence of the two key social triggers (observations of others and advice from others) rather than on encompassing and typifying the triggers. Future work should explore a variety of behavioral triggers for risky behaviors through observation and in-depth interviews. Additionally, we asked participants about the representative risky behaviors that would be expected to be prevalent among users. It is important to note that the risky behaviors we asked about do not always pose a security and privacy threat to users. In the future, we need to explore the triggers according to the risk levels of risky behaviors.

Second, people may generally be reluctant to report that they have engaged in risky behaviors, and thus, responses are subject to social desirability bias. To reduce this bias, we told the participants at the beginning of the survey that we did not consider engaging in risky behaviors to be bad or uncommon and requested honest and accurate responses.

Recruitment of participants. Because Prolific workers have more technology knowledge than the general U.S. population [2, 55], the percentage of users who engage in risky behavior may be higher than the results of this study.

While Das et. al [11] recruited participants mainly from the U.S. and India and found cultural differences (people from India were significant more likely to report social triggers for S&P behaviors), we decided to recruit participants only in the U.S. We initially considered conducting this study in Japan, which has the lowest SeBIS score [51]. We recruited Japanese participants through Lancers [33], a popular crowdsourcing platform in Japan. We found that the percentage of Japanese participants who reported engaging in risky behavior was much lower than that of the U.S. participants (e.g., public Wi-Fi: 3.2% in Japan, 16.3% in the U.S.). Because Lancers, unlike Prolific, is not academic-specific and is used for a variety of tasks including data analysis, it might have more technically skilled workers than Prolific. Researchers can reach Indian workers through MTurk, but the data quality is generally lower than Prolific [38, 55]. Another reason we did not compare the U.S. to other countries is that the U.S. has been treated as an individualistic country in the past, but the most recent Hofstede's individualism score of the U.S. is much lower than it used to be (from 91 to 60, updated in October 2023) [24]. In the future, we will need to compare a country with a much higher individualism score than the U.S. with a country with a much lower score.

4 Results

We present the survey results to address our research questions: the frequency of social triggers for the risky behaviors (RQ1), the source of social triggers (RQ2), the factors of the social triggers (RQ3), and user practices in sharing the risky behaviors (RQ4).

4.1 RQ1: Frequency of Social Triggers

Table 2 shows the percentage of participants who engaged in each risky behavior and the frequency of behavioral triggers that led to each behavior.

Risky behaviors. The frequently reported risky behaviors were password reuse and delayed update, with 71.2% (297/417) of our participants reporting having reused the same or similar passwords and 61.2% (255/417) reporting having ignored or delayed software/application updates in the 6 months preceding the survey. Additionally, 32.4% (135/417) reported having shared an account with others, 18.2% (76/417) reported having downloaded illegal or unofficial content, 18.2% (76/417) reported having shared their sensitive personal information online, and 16.3% (68/417) reported having connected to public Wi-Fi. Overall, 90.9% (379/417) of our participants reported having engaged in at least one of the six risky behaviors. This result suggests that risky behaviors are common among users and that S&P researchers should work to reduce such user practices.

Behavioral triggers. As shown in Table 2, 23.8% of the participants had observed/heard about others engaging in risky behavior before engaging in the behavior on average. On the other hand, fewer participants had experienced advice from others or coercion from an organization; 6.8% had been advised and 3.4% had been required to engage in risky behavior. The majority (70.0%) of the participants had not experienced any of the three triggers above.

We found that the frequency of the triggers, especially social vs. voluntary triggers, varied depending on the type of risky behaviors. While many participants had voluntarily reused passwords and delayed updates, about half of participants observed others or received advice from others who downloaded illegal content and shared accounts. We explain statistical differences in the frequency of social triggers among risky behaviors in Section 4.3.

Although the majority of participants engaged in risky behaviors solely of their own volition, we cannot ignore the fact that about one third of participants were influenced by social triggers to reduce risky user behaviors. We cannot measure which of the user voluntary volition or social triggers had a greater impact on participants' decisions to engage in risky behaviors as discussed in Section 3.4, but the approach of reducing negative social effects could be helpful in reducing risky user behaviors.

Table 2: Frequency of behavioral triggers for risky behaviors.

Behavioral triggers		Public Wi-Fi N=68 (16.3%)	Pwd Reuse N=297 (71.2%)	Illegal DL N=76 (18.2%)	Delayed Update N=255 (61.2%)	Sensitive Post N=76 (18.2%)	Account Sharing N=135 (32.4%)	All
Social	Observation	33.8%	17.5%	48.7%	14.1%	26.3%	35.6%	23.8%
	Advice	4.4%	1.3%	13.2%	5.5%	10.5%	17.0%	6.8%
Organizational		8.8%	3.0%	1.3%	2.7%	6.6%	2.2%	3.4%
Voluntary		61.8%	79.8%	47.4%	78.8%	63.2%	52.6%	70.0%

We show the percentage of the participants who had experienced each trigger (observation of others' behavior, advice from others, or organizational enforcement) among the participants who reported having engaged in each risky behavior. Participants could select more than one trigger. Therefore, the sum of each column exceeds 100%. On the other hand, a voluntary trigger means the participant had not experienced any of the three triggers above (i.e., an exclusive option).

Table 3: Person engaging in risky behaviors that participants had observed/heard about.

	Public Wi-Fi (N=23)	Pwd Reuse (N=52)	Illegal DL (N=37)	Delayed Update (N=36)	Sensitive Post (N=20)	Account Sharing (N=48)	All
Friend	**65.2%**	**65.4%**	**70.3%**	47.2%	**80.0%**	**64.6%**	**64.4%**
Family	39.1%	**55.8%**	16.2%	38.9%	**55.0%**	**77.1%**	49.1%
Stranger/Online posts	30.4%	34.6%	**67.6%**	36.1%	40.0%	25.0%	38.4%
Colleague	**52.2%**	26.9%	18.9%	27.8%	20.0%	20.8%	26.4%
Media (e.g., news and TV programs)	21.7%	13.5%	10.8%	11.1%	15.0%	12.5%	13.4%
Influencer	13.0%	5.8%	10.8%	8.3%	25.0%	8.3%	10.2%
Teacher/Mentor	4.3%	3.8%	2.7%	0.0%	5.0%	4.2%	3.2%

The first row shows the number of the participants who had observed others engaging in each risky behavior before engaging in the behavior. The sum of the percentages for each behavior exceeds 100% because the participants could select more than one type of relationship. Bold numbers highlight items greater than 50%.

4.2 RQ2: Source of Social Triggers

Existing studies have showed that non-expert users have various sources of security advice, such as family, friends, colleagues, and technical support [11, 45, 48, 49]. We were interested in from whom users learn about *risky* behaviors.

Table 3 shows the person engaging in risky behaviors that participants had observed/heard about. For the five risky behaviors other than delayed update, more than 60% of participants had observed/heard about their friend engaging in the behavior. In particular, 80.0% of participants had observed/heard about their friend sharing sensitive personal information online to strangers on social media. Family was the second most common, with an average of about half (49.1%) of participants having observed/heard about risky behaviors of their family members and especially 77.1% having observed/heard about account sharing.

We found that participants had observed/heard about the risky behaviors of online strangers relatively frequently. In particular, 67.6% of participants had encountered strangers downloading illegal or unofficial content. The result indicates that the social triggers for risky behaviors occur both offline and online.

More than half (52.2%) of participants had seen or heard about their colleague connecting to public Wi-Fi. Some participants had observed/heard about risky behaviors from media (e.g., news websites and TV programs) and influencers, while few participants had observed/heard about teachers or mentors.

We show who advised the participants to engage in risky behaviors in Table 9 of Appendix B. Please note that the number of others advising the participants was less than the number of others being observed by the participants.

We were interested in whether the person by whom users are triggered differs by user demographics. Table 4 shows which relationships led to socially-triggered risky behaviors by participant demographics. We performed Fisher's exact tests to test whether the proportions differed by user group (p-values were adjusted using the Bonferroni method). We found that, for friends, family, and colleagues, there were no significant differences in the proportions across user groups. On the other hand, there were significant differences in the proportion of risky behaviors triggered by online strangers across the age groups of the participants. Specifically, younger participants' socially-triggered risky behaviors were more likely to be triggered by online strangers ($p < 0.001$ for the 18–34 age group (45.3%) vs. the 60+ age group (13.3%);

Table 4: Relationships between participant demographics and those who influenced them.

		Friend	Family	Stranger	Colleague
	18–34	71.6%	50.5%	**45.3%**	30.5%
Age	35–59	51.6%	42.2%	**39.1%**	23.4%
	≥ 60	64.4%	57.8%	**13.3%**	24.4%
Male		62.9%	47.4%	37.9%	31.9%
Female		64.8%	52.3%	34.1%	20.5%
	6–14	60.9%	52.2%	52.2%	13.0%
SA-6	15–24	63.2%	54.4%	36.0%	28.8%
	25–30	66.1%	37.5%	30.4%	28.6%

We show the proportions of risky behaviors triggered by a particular relationship to those triggered by others for each user group. Note that this does not show how each user group relates to the likelihood of risky behaviors or the likelihood of social triggers (which we show in Table 5). Bold text indicates that there was a significant difference in the proportions.

p =0.014 for the 35–59 age group (39.1%) vs. the 60+ age group (13.3%)). Given that young people generally spend more time online [53], it is perhaps not surprising that they are more likely to observe risky behaviors of online strangers.

4.3 RQ3: Factors of Social Triggers

To understand the factors of social triggers for risky user behaviors, we performed a logistic regression. Specifically, we modeled how likely a participant would be to report a social trigger given their age, gender, SA-6, and the type of risky behavior they reported having engaged in. We used random intercepts for each participant to consider repeated observations. We calculated fifteen pairwise comparisons between the six different risky behaviors using R's multcomp package [25]. We corrected the significance levels due to the multiple comparisons using the Bonferroni method [1]. In addition, we ran an ordinal logistic regression to understand the demographics of users who are generally more likely to engage in risky behaviors, regardless of the trigger type. The dependent variable was the number of risky behaviors that a participant reported engaging in. Table 5 shows the results of the two logistic regressions. A positive coefficient implies that the independent variable has a positive effect on the dependent variable, while a negative coefficient implies the opposite. Coefficients imply the expected change in log odds of having the outcome per unit change in the independent variable. More specifically, the odds ratio (*OR*) indicates the change in the odds of the outcome (e.g., odds of how likely participants report a social trigger) for a 1-unit increase in the continuous independent variable (e.g., 1-score increase of participants' SA-6) or compared with a reference categorical independent variable (e.g., male participants compared with female participants).

Individual demographics. We found that while individual demographics were significantly correlated with risky behaviors,

Table 5: Logistic regressions for risky behaviors and social triggers (coefficients and *p*-values).

	Social Triggers		Risky Behaviors	
Age	−0.001		−0.050	***
Male (vs. Female)	0.334		0.420	*
SA-6	0.068	*	−0.099	***
DL (vs. Pwd)	2.503	***		
Account (vs. Pwd)	2.139	***		
Update (vs. DL)	−2.424	***		
Account (vs. Update)	2.060	***		

The middle column shows the results of a logistic regression explaining whether social triggers had occurred before participants engaged in risky behaviors. Of the fifteen pairwise comparisons of risky behaviors, only those pairs with a significant difference are shown in this table. The right column shows the results of an ordinal logistic regression explaining the number of risky behaviors reported by participants. Significance levels: *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

they were less correlated with whether the risky behaviors were socially triggered.

In terms of age and gender, younger participants were significantly more likely to engage in the risky behaviors we examined in this study (*coeff* = −0.050, *OR* = 0.951, $p < 0.001$) and male participants were significantly more likely to engage in the risky behaviors (*coeff* = 0.420, *OR* = 1.522, $p = 0.020$). These results are consistent with the study by Milne et al. [42], which concluded that younger and male online shoppers in the U.S. were more likely to adopt risky behaviors. On the other hand, we found no significant correlations between age and the likelihood of social triggers (*coeff* = −0.001, *OR* = 0.999, $p = 0.917$) and between gender and the likelihood of social triggers (*coeff* = 0.334, *OR* = 1.396, $p = 0.292$). Das et al. [11] found that younger people were more likely to report social triggers for S&P behaviors, but gender was not correlated with this, and then suggested that some level of age-based personalization may be needed to trigger user S&P behaviors. Such age-based personalization may be effective in socially promoting S&P behaviors but may be less effective in reducing socially triggered risky behaviors.

In terms of security attitude (SA-6), we found that participants with a lower SA-6 score were significantly more likely to engage in the risky behaviors we examined in this study (*coeff* = −0.099, *OR* = 0.906, $p < 0.001$), which is consistent with our intuition. On the other hand, participants with a higher SA-6 score were significantly more likely to report social triggers (*coeff* = 0.068, *OR* = 1.071, $p = 0.030$). It may be possible that users with high security attitudes are less likely to engage in risky behaviors for voluntary motivations such as convenience, but they may think it would be okay to engage in the behaviors if they observe others engaging in them. Note, however, that we did not collect the data to conclude that users with high security attitudes had not observed
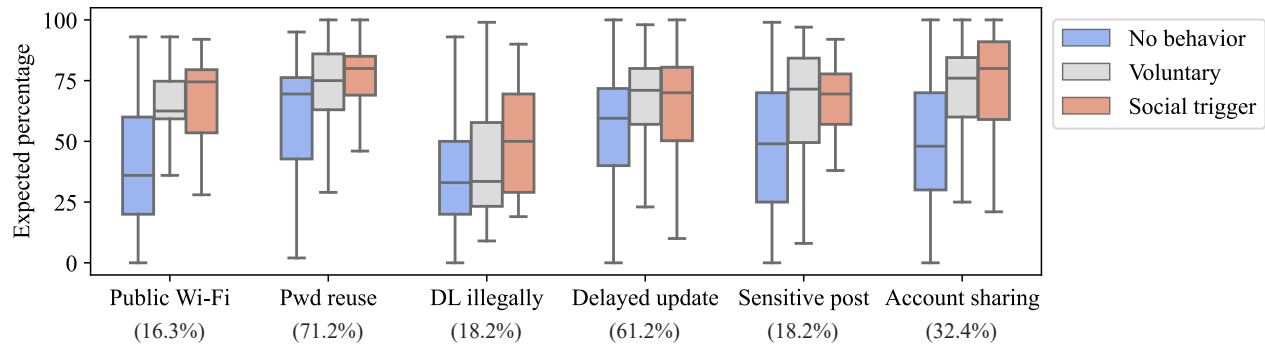
Figure 2: Expected percentages of public engaging in risky behavior.
The numbers in parentheses indicate the percentage of the participants who reported engaging in the behavior.

risky behaviors of others when they decided not to engage in the behaviors, and thus, we need further investigation.

Risky behaviors. In contrast to individual demographics, the type of risky behaviors was significantly correlated with whether the risky behaviors were socially triggered, i.e., several risky behaviors were more likely to be socially triggered. Our regression analysis revealed significant differences across risky behaviors controlling for age, gender, and SA-6, as shown in Table 5. Of the fifteen pairwise of comparisons of risky behaviors, we found significant differences in four as follows. Illegal downloading was significantly more likely to have reported social triggers than password reuse (*coeff* = 2.503, *OR* = 12.221, *p* < 0.001). Delayed update was significantly less likely to have reported social triggers than illegal downloading (*coeff* = −2.424, *OR* = 0.089, *p* < 0.001). Account sharing was significantly more likely to have reported social triggers than password reuse (*coeff* = 2.139, *OR* = 8.493, *p* < 0.001) and delayed update (*coeff* = 2.060, *OR* = 7.843, *p* < 0.001). In summary, illegal downloading and account sharing are more likely to be socially triggered, as opposed to password reuse and delayed update.

Risk information. We asked participants having engaged in risky behaviors due to advice from others whether they had been told about the risks of the behavior by the person. We found that they were not always told about the risks; of the reported risky behaviors that were triggered by advice from others, 59.6% of the behaviors occurred when the participants had not been told about the risks.

Expected risky behaviors of others. We were also interested in what the participants who had engaged in the risky behavior, especially those who had experienced social triggers (i.e., observations of others or advice from others), expected the percentage of the public who engaged in risky behavior to be. This could provide insights into how users generalize that they have observed risky behaviors of their friends and family and how they generalize their own risky behavior. Note that our data would only show correlation, not causation, i.e., we

cannot conclude that users engage in risky behaviors as a result of their expectations that most of the public engage in the behaviors.

Figure 2 shows box plots indicating the expected percentage of the public who engage in each risky behavior among three groups: the participants who did not engage in the behavior, those who voluntarily engaged in the behavior, and those who experienced social triggers. Due to the non-normal distribution, we compared the three groups by using Kruskal-Wallis tests and then compared each pair by using post-hoc Steel-Dwass tests. The significance levels were corrected using the Bonferroni method for multiple comparisons [1]. For all risky behaviors, the median of the no-behavior group was lower than that of the other two groups, and the differences were significant for all pairs except for the voluntary group for illegal downloading. In other words, those who engaged in risky behaviors tended to expect more of the public to engage in the behavior than those who did not. On the other hand, there was no significant difference between the voluntary and social-trigger groups, possibly due to the small sample size of the social-trigger group. This does not mean differences do not exist but rather that they might be too slight to detect at smaller sample sizes. When considering the medians instead of just *p*-values, we found that the social-trigger group had a higher median than the voluntary group for 4 out of the 6 risky behaviors.

We also found that all of the medians of the expected percentages were higher than the percentage of our participants who reported engaging in the risky behaviors (e.g., 16.3% of our participants connected to public Wi-Fi), except for the non-behavior group for illegal downloading. This may be somewhat natural given that participants from Prolific have more technology knowledge than the general U.S. population [2, 55]

The fact that users engaging in risky behaviors tend to expect more of the public to engage in the behaviors may contribute to the users continuing practices of the risky behaviors, even if the expectation may not be their initial motivation.

Table 6: Participants' sharing practices and person with whom they shared.

	Public Wi-Fi (N=68)	Pwd Reuse (N=297)	Illegal DL (N=76)	Delayed Update (N=255)	Sensitive Post (N=76)	Account Sharing (N=135)	All
Overall shared	26.5%	24.2%	52.6%	27.2%	46.1%	80.7%	37.8%
Family	**55.6%**	**70.8%**	40.0%	**53.6%**	48.6%	**86.2%**	**65.6%**
Friend	50.0%	45.8%	**77.5%**	39.1%	**51.4%**	47.7%	49.6%
Colleague	38.9%	6.9%	5.0%	21.7%	11.4%	7.3%	12.0%
Online discussion	11.1%	2.8%	10.0%	11.6%	28.6%	1.8%	8.2%

The first row indicates the number of participants who reported engaging in each risky behavior, and the second row indicates the percentage of participants who shared the behavior with others among those who reported engaging in the behavior. The third and subsequent rows indicate the percentage of participants who shared the risky behavior with a specific person among those who shared the behavior. The sum of the percentages for each behavior exceeds 100% because the participants could select more than one type of relationship. Bold numbers highlight items greater than 50%.

4.4 RQ4: Sharing Practices

Frequency of sharing. Das et al. [11] found that 32% of S&P behaviors were shared with others. This suggests a promising phenomenon of stories about S&P practices being widespread among users. We show the frequency of sharing *risky* behaviors in Table 6. On average, 37.8% of risky behaviors were shared with others, although the frequency of sharing varied considerably by behavior type (see Table 8 for the regression analysis). This means that stories about risky behaviors are spreading among users as much or more than stories about S&P practices.

Table 6 also shows the person with whom participants shared their risky behaviors. Just as participants often observed their friends and family members engaging in risky behaviors (as shown in Table 3), they often shared their risky behaviors with family and friends. On the other hand, it is interesting to note that while participants often observed online strangers engaging in risky behaviors, they seldom shared their risky behaviors with strangers on online discussion sites. This asymmetry suggests a large impact relative to the number of people sharing risky behaviors online, i.e., one person's post about risky behaviors could be seen by many users.

Reasons for sharing. Table 7 shows the reasons why participants shared their risky behaviors with others. The most common reason was "I wanted them to get the benefits." Naturally, users do not share their risky behaviors with others for malicious purposes; rather, they simply want others to get the benefits, such as convenience. The second most common reason was "I just wanted to talk about my recent behavior," which Das et al. [11] found to be the most common reason for sharing S&P behaviors. The third most common reason, "I wanted them to know about other options, regardless of risk," also indicates that the participants valued other objectives, such as convenience, more than the risk of the behavior. The

participants who selected "They noticed my change" may not have initially had a clear intention to share their risky behaviors.

The reasons given by participants in open-ended form as "other" include "to share a complaint" (e.g., a participant answered "I complained that I am sick of these forced <OS name> updates so frequently so I put them off") and "Others confided in me first" (e.g., "They told me they do this").

We also asked participants who reported engaging in risky behaviors but did not share their behaviors about their reasons for doing so. The primary reasons were "I just didn't want to talk about this with anyone" (54.3%) and "I assumed everyone did this" (33.3%).

Factors of sharing. To understand the factors of sharing practices, we performed a logistic regression modeling how likely a participant was to share their risky behavior given their age, gender, SA-6, the type of risky behaviors, and whether their behavior was socially triggered. In the same way as Table 5, we used random intercepts for each participant to consider repeated observations and calculated the fifteen pairwise comparisons between the six different risky behaviors using R's multcomp package [25]. We corrected the significance levels using the Bonferroni method [1]. Table 8 shows the result.

Das et al. [11] found no significant correlations between user sharing practices of S&P behaviors and individual demographics (age, gender, and SeBIS). We also found no significant correlation between user sharing practices of *risky* behaviors and individual demographics (age, gender, and SA-6).

On the other hand, we found significant correlations between user sharing practices and the type of risky behaviors. Specifically, as shown in Table 8, all pairwise differences between the likelihood of sharing practices of account sharing and each of the other behaviors were significant. From the results of Table 6, next to account sharing, illegal downloading was likely to be shared, followed by sensitive posts.

Importantly, we also found a significant correlation between user sharing practices and whether their behavior was

Table 7: Reasons for sharing risky behaviors with others.

	Public Wi-Fi (N=18)	Pwd Reuse (N=72)	Illegal DL (N=40)	Delayed Update (N=69)	Sensitive Post (N=35)	Account Sharing (N=109)	All
I wanted them to get the benefits	50.0%	28.8%	45.0%	10.0%	25.7%	**73.4%**	41.7%
I just wanted to talk about my recent behavior	33.3%	45.2%	42.5%	45.7%	**57.1%**	14.7%	35.9%
I wanted them to know about other options	5.6%	16.4%	42.5%	15.7%	5.7%	5.5%	14.2%
They noticed my change	27.8%	11.0%	0.0%	12.9%	17.1%	11.9%	11.9%
I wanted them to know about my knowledge	22.2%	2.7%	2.5%	2.9%	5.7%	1.8%	3.8%
Other	5.6%	12.3%	12.5%	15.7%	11.4%	5.5%	10.4%

The first row indicates the number of participants who shared their risky behavior with others. The sum of the percentages for each behavior exceeds 100% because the participants could select more than one reason. Bold numbers highlight items greater than 50%.

Table 8: Logistic regression for sharing practices.

	$Coeff$	p-value	
Age	0.004	0.553	
Male (vs. Female)	−0.231	0.327	
SA-6	0.024	0.289	
Account (vs. Wi-Fi)	2.794	< 0.001	***
Account (vs. Pwd)	2.833	< 0.001	***
Account (vs. DL)	1.684	0.011	*
Account (vs. Update)	2.696	< 0.001	***
Account (vs. Post)	1.997	0.001	**
Social Trigger	2.062	< 0.001	***

Of the fifteen pairwise comparisons of risky behaviors, only those pairs with a significant difference are shown in this table. Significance levels: $*p < 0.05$; $**p < 0.01$; $***p < 0.001$.

socially triggered ($coeff$ = 2.062, OR = 7.862, $p < 0.001$). In other words, if a participant engaged in risky behavior due to social triggers, they were more likely to share it with others. Specifically, while 25.9% of the risky behaviors caused by other triggers (organizational or voluntary) were shared with others, 70.7% of the risky behaviors caused by social triggers were shared (i.e., 2.7 times more frequently).

5 Discussions

5.1 Summary of Findings

- Social triggers can lead users to engage in risky behaviors. Specifically, approximately 20–50% of the participants observed others engaging in risky behavior or were advised to do so before engaging in the risky behavior. (RQ1)

- Risky user behaviors are primarily triggered by friends, family, and online strangers. (RQ2)

- The likelihood of social triggers is significantly correlated with the type of risky behavior. In other words, certain behaviors (account sharing and illegal downloading) are often caused by social triggers. (RQ3)

- Risky behaviors caused by social triggers are more likely to be shared with others (i.e., negative social chains). Users share their risky behaviors primarily because they want others to get the benefit. (RQ4)

5.2 Recommendations

We showed that participants engaged in risky behaviors following observations of others and/or advice from others. Researchers should work to reduce such negative effects of social triggers, but this issue is not so straightforward. In extreme cases, preventing users from having social connections would protect them from negative social effects, but this is an impractical measure. Most importantly, social triggers also have positive effects. As Das et al.'s study [11] and the other previous studies we mentioned in Section 2.2 demonstrated, users often engage in S&P behaviors due to social triggers, such as receiving security advice from others. Therefore, researchers need to work simultaneously on activating the positive aspects of social triggers and reducing the negative aspects.

It is also important to note that it is essential for researchers to work to reduce the number of users engaging in risky behaviors voluntarily, as our results show that the majority of participants engaged in risky behaviors voluntarily. For this purpose, basic security education and the interventions that have been proposed in the usable privacy and security field, such as nudges and warnings [14], would be effective.

In the following, we suggest approaches to reducing the *negative* aspects of social triggers (i.e., triggering *risky* user behaviors).

Interventions on online platforms. We found that participants were triggered to engage in risky behaviors not only by offline social connections, such as friends, family, and colleagues, but also by online strangers and influencers. In particular, downloading of illegal or unofficial content was often triggered by online strangers. This suggests the need for interventions to combat the negative chains of risky behaviors that occur online. Online intervention is important because a single post about risky behaviors can be seen by

multiple users, meaning that negative social chains can be easily amplified. Online intervention would be less difficult than eliminating of offline negative chains. Specifically, we recommend that online platforms formulate or strengthen their guidelines regarding posts encouraging risky behaviors and reporting risky behaviors, detect such posts, and present warnings for such posts. Our work would help online platforms identify the risky behaviors for which they should implement the above interventions. In the field of dis/misinformation research, researchers have evaluated the effective design of warnings to prevent the spread of dis/misinformation [22, 29]. The findings of those studies may also be useful in preventing the spread of risky behaviors online. In addition to direct interventions by online platforms, we suggest that online platforms provide features for S&P experts or the public to intervene, such as reporting or correction features. The above efforts should be made not only for posts encouraging illegal downloading (at the request of the copyright owners of the original content) but for any risky behavior susceptible to negative social chains.

Security education with emphasis on risky behaviors susceptible to negative social chains and the risks. We show that the type of risky behaviors is more likely to influence the likelihood of social triggers for risky behaviors and practices of sharing risky behaviors than individual demographics. This suggests the need for countermeasures specific to risky behaviors that are prone to negative social chains, rather than personalized countermeasures tailored to individual demographics. Incorporating such risky behaviors into security education materials or conducting activities to publicize the risks of such risky behaviors would be an effective way to combat negative social chains. In addition, we showed that participants shared their risky behaviors with others because they wanted others to get the benefits or to know about other options, regardless of the risks. In other words, users share their risky behaviors in favor of benefits (e.g., convenience) rather than risks. On the basis of this fact, we recommend that security education not only introduce non-recommended risky behaviors but also convey the risks together. Risk information should be conveyed in an impressive way that is easy for users to understand and remember, such as by quantifying the degree of risks and showing the risks in a graphic or video presentation. For example, graphic cigarette packages that depict the risks of smoking have successfully reduced the demand for cigarettes [57]. For another example, exposure to a drama that focused on the aversive consequences of traffic accidents successfully raised people's awareness of the potentially negative consequences of traffic accidents [44]. In the S&P field, researchers have already worked on visualizations of specific types of risks (e.g., unsafety of URLs [6] and data collection by IoT devices [17]), but it would be desirable to propose and evaluate designs for visualizing the risks of diverse risky behaviors.

Removal of expectation that most of public engages in risky behaviors. We found that participants engaging in a risky behavior expected a higher percentage of the public to engage in the risky behavior than those who did not and that it is possible that participants engaging in a risky behavior due to social triggers expected an even higher percentage. We cannot conclude that such expectations are a cause of risky user behaviors, but dispelling such expectations would be effective in preventing users from continuing to engage in risky behaviors. Interventions that present the percentage of security experts who would not engage in a risky behavior before a user engages in the behavior may be effective in dispelling such user expectations. As related interventions, interventions for dispelling user expectations of others with respect to the phenomenon of pluralistic ignorance (i.e., users do not really want to do something but do it because they think everyone else is doing it) have been proposed and discussed [47, 52]. For example, holding discussions to learn about the true beliefs of others was effective in dispelling user expectations of others [52]. However, it may be impractical to apply those interventions for dispelling user expectations of the risky behaviors of others because risky behaviors often bring users benefits.

6   Conclusion and Future Work

To improve user security and privacy behaviors, researchers need to not only focus on the attitudes and behaviors of individual users but also understand the relationship between each user and society. We analyzed the effects of social triggers for risky behaviors and found that participants sometimes engaged in risky behaviors after observing others or getting advice from others. Participants shared their risky behaviors with others primarily to let others get the benefits.

Future work should examine behavioral triggers for more diverse risky behaviors in multiple countries/cultures, especially individualistic and collectivistic countries. In addition, we need to implement interventions to reduce the negative social effects and evaluate their effectiveness in the future.

References

[1] Hervé Abdi. Bonferonni and šidák corrections for multiple comparisons. *Encyclopedia of measurement and statistics*, 3(01), 2007.

[2] Desiree Abrokwa, Shruti Das, Omer Akgul, and Michelle L. Mazurek. Comparing security and privacy attitudes among U.S. users of different smartphone and smart-speaker platforms. In *Proceedings of the 17th Symposium on Usable Privacy and Security*, SOUPS'21, 2021.

[3] Ruba Abu-Salma, M. Angela Sasse, Joseph Bonneau, Anastasia Danilova, Alena Naiakshina, and Matthew Smith. Obstacles to the adoption of secure communication tools. In *Proceedings of the 2017 IEEE Symposium on Security and Privacy*, S&P'17, 2017.

[4] Omer Akgul, Richard Roberts, Moses Namara, Dave Levin, and Michelle L. Mazurek. Investigating influencer vpn ads on youtube. In *Proceedings of the 2022 IEEE Symposium on Security and Privacy*, S&P'22, 2022.

[5] Suzan Ali, Tousif Osman, Mohammad Mannan, and Amr Youssef. On privacy risks of public WiFi captive portals. In *Proceedings of the Data Privacy Management, Cryptocurrencies and Blockchain Technology: ESORICS 2019 International Workshops*, DPM CBT'19, 2019.

[6] Kholoud Althobaiti, Nicole Meng, and Kami Vaniea. I don't need an expert! making url phishing features human comprehensible. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI'21, 2021.

[7] Daniel V. Bailey, Collins W. Munyendo, Hunter A. Dyer, Miles Grant, Philipp Markert, and Adam J Aviv. "someone definitely used 0000": Strategies, performance, and user perception of novice smartphone-unlock pin-guessers. In *Proceedings of the 2023 European Symposium on Usable Security*, EuroUSEC'23, pages 158–174, 2023.

[8] Sruti Bhagavatula, Lujo Bauer, and Apu Kapadia. "Adulthood is trying each of the same six passwords that you use for everything": The scarcity and ambiguity of security advice on social media. In *Proceedings of the 25th ACM Conference on Computer-Supported Cooperative Work and Social Computing*, CSCW'22, 2022.

[9] Karoline Busse, Julia Schäfer, and Matthew Smith. Replication: No one can hack my mind revisiting a study on expert and non-expert security practices and advice. In *Proceedings of the 15th Symposium on Usable Privacy and Security*, SOUPS'19, 2019.

[10] Anupam Das, Joseph Bonneau, Matthew Caesar, Nikita Borisov, and XiaoFeng Wang. The tangled web of password reuse. In *Proceedings of the 2014 Network and Distributed System Security Symposium*, NDSS'14, 2014.

[11] Sauvik Das, Laura A. Dabbish, and Jason I. Hong. A typology of perceived triggers for end-user security and privacy behaviors. In *Proceedings of the 15th Symposium on Usable Privacy and Security*, SOUPS'19, 2019.

[12] Sauvik Das, Adam D.I. Kramer, Laura A. Dabbish, and Jason I. Hong. Increasing security sensitivity with social proof: A large-scale experimental confirmation. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, CCS'14, 2014.

[13] Sauvik Das, Adam D.I. Kramer, Laura A. Dabbish, and Jason I Hong. The role of social influence in security feature adoption. In *Proceedings of the 18th ACM conference on Computer Supported Cooperative Work & Social Computing*, CSCW'15, 2015.

[14] Verena Distler, Gabriele Lenzini, Carine Lallemand, and Vincent Koenig. The framework of security-enhancing friction: How ux can help users behave more securely. In *Proceedings of the New Security Paradigms Workshop 2020*, NSPW'20, 2020.

[15] Daniel Domínguez-Álvarez, Daniel Toniuc, and Alessandra Gorla. Rechan: an automated analysis of android app release notes to report inconsistencies. In *Proceedings of the 9th IEEE/ACM International Conference on Mobile Software Engineering and Systems*, MobileSoft'22, 2022.

[16] Serge Egelman and Eyal Peer. Scaling the security wall: Developing a security behavior intentions scale (SeBIS). In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, CHI'15, 2015.

[17] Pardis Emami-Naeini, Yuvraj Agarwal, Lorrie Faith Cranor, and Hanan Hibshi. Ask the experts: What should be on an iot privacy and security label? In *Proceedings of the 2020 IEEE Symposium on Security and Privacy*, S&P'20, 2020.

[18] Pardis Emami-Naeini, Martin Degeling, Lujo Bauer, Richard Chow, Lorrie Faith Cranor, Mohammad Reza Haghighat, and Heather Patterson. The influence of friends and experts on privacy decision making in iot scenarios. In *Proceedings of the 21st ACM Conference on Computer-Supported Cooperative Work and Social Computing*, CSCW'18, 2018.

[19] Michael Fagan and Mohammad Maifi Hasan Khan. Why do they do what they do?: A study of what motivates users to (not) follow computer security advice. In *Proceedings of the 12th symposium on usable privacy and security*, SOUPS'16, 2016.

[20] Cori Faklaris, Laura A. Dabbish, and Jason I. Hong. A self-report measure of end-user security attitudes (SA-6). In *Proceedings of the 15th Symposium on Usable Privacy and Security*, SOUPS'19, 2019.

[21] Brian J. Fogg. A behavior model for persuasive design. In *Proceedings of the 4th international Conference on Persuasive Technology*, PT'09, 2009.

[22] Katrin Hartwig, Frederic Doell, and Christian Reuter. The landscape of user-centered misinformation interventions–a systematic literature review. *arXiv preprint arXiv:2301.06517*, 2023.

[23] Ayako A. Hasegawa, Naomi Yamashita, Tatsuya Mori, Daisuke Inoue, and Mitsuaki Akiyama. Understanding non-experts security-and privacy-related questions on a Q&A site. In *Proceedings of the 18th Symposium on Usable Privacy and Security*, SOUPS'22, 2022.

[24] Hofstede Insights. Country comparison tool. `https://www.hofstede-insights.com/country-comparison-tool`, (accessed January 15, 2024).

[25] Torsten Hothorn, Frank Bretz, Peter Westfall, Richard M. Heiberger, Andre Schuetzenmeister, and Susan Scheibe. multcomp: Simultaneous inference in general parametric models. `https://cran.r-project.org/web/packages/multcomp/index.html`, (accessed February 8, 2024).

[26] Iulia Ion, Rob Reeder, and Sunny Consolvo. "... no one can hack my mind"': Comparing expert and non-expert security practices. In *Proceedings of the 11th Symposium On Usable Privacy and Security*, SOUPS'15, 2015.

[27] Ipsos. Ipsos global trends report 2023. `https://www.ipsos.com/en/global-trends`, (accessed December 29, 2023).

[28] Shareen Irshad and Tariq Rahim Soomro. Identity theft and social media. *International Journal of Computer Science and Network Security*, 18(1):43–55, 2018.

[29] Ben Kaiser, Jerry Wei, Eli Lucherini, Kevin Lee, J. Nathan Matias, and Jonathan Mayer. Adapting security warnings to counter online disinformation. In *Proceedings of the 30th USENIX Security Symposium*, SEC'21, pages 1163–1180, 2021.

[30] kaspersky. Stranger danger: the connection between sharing online and losing the data we love. `https://www.kaspersky.com/blog/my-precious-data-report-three/16883/`, (accessed December 29, 2023).

[31] Jess Kropczynski, Reza Ghaiumy Anaraky, Mamtaj Akter, Amy J. Godfrey, Heather Lipford, and Pamela J. Wisniewski. Examining collaborative support for privacy and security in the broader context of tech caregiving. In *Proceedings of the 24th ACM Conference on Computer-Supported Cooperative Work and Social Computing*, CSCW'21, 2021.

[32] Isadora Krsek, Kimi Wenzel, Sauvik Das, Jason I. Hong, and Laura Dabbish. To self-persuade or be persuaded: Examining interventions for users' privacy setting selection. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI'22, 2022.

[33] Lancers, Inc. Lancers. `https://www.lancers.jp/`, (accessed December 12, 2023).

[34] Alice Yuen Loke and Yim-wah Mak. Family process and peer influences on substance use by adolescents. *International journal of environmental research and public health*, 10(9):3868–3885, 2013.

[35] Philipp Markert, Daniel V Bailey, Maximilian Golla, Markus Dürmuth, and Adam J Aviv. On the security of smartphone unlock pins. *ACM Transactions on Privacy and Security (TOPS)*, 24(4):1–36, 2021.

[36] Hiroaki Masaki, Kengo Shibata, Shui Hoshino, Takahiro Ishihama, Nagayuki Saito, and Koji Yatani. Exploring nudge designs to help adolescent sns users avoid privacy and safety threats. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI'20, 2020.

[37] Arunesh Mathur, Josefine Engel, Sonam Sobti, Victoria Chang, and Marshini Chetty. "they keep coming back like zombies": Improving software updating interfaces. In *Proceedings of the 12th Symposium on Usable Privacy and Security*, SOUPS'16, 2016.

[38] Tenga Matsuura, Ayako A. Hasegawa, Mitsuaki Akiyama, and Tatsuya Mori. Careless participants are essential for our phishing study: Understanding the impact of screening methods. In *Proceedings of the 2021 European Symposium on Usable Security*, EuroUSEC'21, 2021.

[39] Tara Matthews, Kerwell Liao, Anna Turner, Marianne Berkovich, Robert Reeder, and Sunny Consolvo. "she'll just grab any device that's closer": A study of everyday device & account sharing in households. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI'16, 2016.

[40] Tamir Mendel and Eran Toch. Social support for mobile security: Comparing close connections and community volunteers in a field experiment. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI'23, 2023.

[41] Marina Micheli, Elissa M. Redmiles, and Eszter Hargittai. Help wanted: Young adults' sources of support for questions about digital media. *Information, Communication & Society*, 23(11):1655–1672, 2020.

[42] George R. Milne, Lauren I. Labrecque, and Cory Cromer. Toward an understanding of the online consumer's risky behavior and protection practices. *Journal of Consumer Affairs*, 43(3):449–473, 2009.

[43] Jonathan A. Obar and Anne Oeldorf-Hirsch. The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society*, 23(1):128–147, 2020.

[44] G. O'Brien, F. Rooney, Colm Carey, and Ray Fuller. Evaluation of the effectiveness of a dramatic presentation on attitudes to road safety. In *Behavioural Research in Road Safety: Twelfth Seminar*, 2002.

[45] Katharina Pfeffer, Alexandra Mai, Edgar Weippl, Emilee Rader, and Katharina Krombholz. Replication: Stories as informal lessons about security. In *Proceedings of the 18th Symposium on Usable Privacy and Security*, SOUPS'22, 2022.

[46] Michael L. Pittaro. Cyber stalking: An analysis of online harassment and intimidation. *International journal of cyber criminology*, 1(2):180–197, 2007.

[47] Emilee Rader. Data privacy and pluralistic ignorance. In *Proceedings of the 19th Symposium on Usable Privacy and Security*, SOUPS'23, 2023.

[48] Emilee Rader, Rick Wash, and Brandon Brooks. Stories as informal lessons about security. In *Proceedings of the 8th Symposium on Usable Privacy and Security*, SOUPS'12, 2012.

[49] Elissa M. Redmiles, Sean Kross, and Michelle L. Mazurek. How I learned to be secure: a census-representative survey of security advice sources and behavior. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS'16, 2016.

[50] Shruti Sannon, Natalya N. Bazarova, and Dan Cosley. Privacy lies: Understanding how, when, and why people lie to protect their privacy in multiple online contexts. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, CHI'18, 2018.

[51] Yukiko Sawaya, Mahmood Sharif, Nicolas Christin, Ayumu Kubota, Akihiro Nakarai, and Akira Yamada. Self-confidence trumps knowledge: A cross-cultural study of security behavior. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI'17, 2017.

[52] Christine M. Schroeder and Deborah A. Prentice. Exposing pluralistic ignorance to reduce alcohol use among college students 1. *Journal of Applied Social Psychology*, 28(23):2150–2180, 1998.

[53] statista. Average daily time spent using the internet by 3rd quarter 2023, by age and gender. https://www.st atista.com/statistics/1378510/daily-time-s pent-online-worldwide-by-age-and-gender/, (accessed February 15, 2024).

[54] Peter Story, Daniel Smullen, Yaxing Yao, Alessandro Acquisti, Lorrie Faith Cranor, Norman Sadeh, and Florian Schaub. Awareness, adoption, and misconceptions of web privacy tools. In *Proceedings of the 21st Privacy Enhancing Technologies Symposium*, PETS'21, 2021.

[55] Jenny Tang, Eleanor Birrell, and Ada Lerner. Replication: How well do my results generalize now? the external validity of online privacy and security surveys. In *Proceedings of the 18th symposium on usable privacy and security*, SOUPS'22, 2022.

[56] Jenny Tang, Hannah Shoemaker, Ada Lerner, and Eleanor Birrell. Defining privacy: How users interpret technical terms in privacy policies. In *Proceedings of the 21st Privacy Enhancing Technologies Symposium*, PETS'21, 2021.

[57] James F. Thrasher, Matthew C. Rousu, David Hammond, Ashley Navarro, and Jay R. Corrigan. Estimating the impact of pictorial health warnings and "plain" cigarette packaging: evidence from experimental auctions among adult smokers in the united states. *Health policy*, 102(1):41–48, 2011.

[58] Rick Wash and Molly M. Cooper. Who provides phishing training? facts, stories, and people like me. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, CHI'18, 2018.

[59] Miranda Wei, Eric Zeng, Tadayoshi Kohno, and Franziska Roesner. Anti-privacy and anti-security advice on TikTok: Case studies of technology-enabled surveillance and control in intimate partner and parent-child relationships. In *Proceedings of the 18th Symposium on Usable Privacy and Security*, SOUPS'22, 2022.

[60] Yuxi Wu, W. Keith Edwards, and Sauvik Das. Sok: Social cybersecurity. In *Proceedings of the 2022 IEEE Symposium on Security and Privacy*, S&P'22, 2022.

[61] Yixin Zou, Kevin Roundy, Acar Tamersoy, Saurabh Shintre, Johann Roturier, and Florian Schaub. Examining the adoption and abandonment of security, privacy, and identity theft protection practices. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI'20, 2020.

Appendix

A Questionnaire

A Study on Online Behaviors

We are conducting a study to explore the impact of behavioral triggers on risky cyber behavior. We don't view taking risky behaviors as bad or uncommon. We are interested in it as researchers. Thus, we kindly request that you provide honest and accurate responses. Your input will be crucial in helping us gain insights into this research area.

The survey will consist of multiple-choice and open-ended questions. We once again request that you provide **honest** and **accurate** responses. Rest assured, your answers will remain entirely confidential and will be anonymous. The aggregated data will be published in an academic paper(s) in a form that does not identify individuals.

The estimated duration for completing the survey is 7 minutes. For your time you will be compensated $1.75 for completing the survey. Please note that participation in this research is entirely voluntary, and you reserve the right to withdraw at any point during the survey without any obligation. Should you have any questions, concerns or comments, feel free to contact the Principal Investigator at <email address>.

By clicking the button below, you acknowledge the following:

- Your participation in the study is voluntary.
- You are 18 years of age or older.
- You are aware that you may choose to terminate your participation at any time for any reason.
- You are a resident of United States of America or Japan

Thank you for your participation in advancing our understanding of cyber behavior. Your valuable input contributes significantly to the success of this research.

- ○ I consent to the above and will voluntarily participate in this survey
- ○ I do not consent to the above and will not participate in this survey

Q01. Have you done any of the following in the past 6 months? Please select all that apply.

- ☐ Connecting to an unknown, potentially unsecured public Wi-Fi, and then engaging in sensitive data exchanges, such as transmitting credit card or password details through this connection
- ☐ Reusing same or similar passwords for different accounts
- ☐ Downloading illegal or unofficial software/applications and media (e.g., videos, music, and games)
- ☐ Ignoring or delaying updating software/applications
- ☐ Sharing sensitive personal information online (e.g., location-based information, real-time activities, and pictures of yourself/others) to strangers on social media
- ☐ Sharing accounts with family, friends or others

☐ None of these apply to me

Q02. <*Asked for each behavior selected by participants in Q01.*>
Did any of the following happen before you took the behavior? Please select all that apply.

- ☐ I observed / heard about other people doing this
- ☐ Other people advised to do this
- ☐ My organization required me to do this
- ☐ Other (Please Specify):
- ☐ Nothing in particular happened

Q02.1 <*Asked for each participant that selected 'I observed / heard about other people doing this' in Q02.*>
You observed/heard people around you doing this. Who did you observe/hear? Please select all that apply.

- ☐ Friend
- ☐ Family
- ☐ Colleague
- ☐ Teacher / Mentor
- ☐ Stranger
- ☐ Influencer
- ☐ Media (e.g., news websites and TV programs)
- ☐ Other (Please Specify)
- ☐ I don't remember

Q02.2 <*Asked for each participant that selected 'Other people advised to do this' in Q02.*>
Who advised you to take this behavior? Please select all that apply.

- ☐ Friend
- ☐ Family
- ☐ Colleague
- ☐ Teacher / Mentor
- ☐ Stranger / Online posts
- ☐ Influencer
- ☐ Media (e.g., news websites and TV programs)
- ☐ Service provider / Salesperson
- ☐ Other (Please Specify)
- ☐ I don't remember

Q02.3 <*Asked for each participant that selected 'Other people advised to do this' in Q02.*>
Did the person who advised you take this behavior share any risks of the behavior?

- ○ Yes
- ○ No
- ○ I don't remember

Q03. <*Asked for each behavior selected by participants in Q01.*>
When taking this behavior did you talk about it with anyone else? Please select all that apply.

- ☐ Friend
- ☐ Family
- ☐ Colleague
- ☐ Online discussion (e.g., Social media, blog posts, forums)
- ☐ Other (Please Specify)
- ☐ I didn't talk about this with anyone

Q03.1 <*Asked for each participant that only selected 'I didn't talk about this with anyone' in Q03.*>
Why did you decide not to talk about this behavior to anyone? Please select all that apply.

- ☐ I didn't feel comfortable talking about security and privacy
- ☐ I assumed everyone did this
- ☐ I just didn't want to talk about this to anyone
- ☐ I hadn't had the chance to talk with anyone about this yet
- ☐ Other (Please Specify)

Q03.2 <*Asked for each participant that did not select 'I didn't talk about this with anyone' in Q03.*>
What prompted you to talk about this behavior with them? Please select all that apply.

- ☐ They noticed my change
- ☐ I wanted them to get the benefits
- ☐ I just wanted to talk about my recent behavior
- ☐ I wanted them to know that I have knowledge in the hacking field
- ☐ I wanted them to know about other options, regardless of risk
- ☐ Other (Please Specify)

Q04. What percentage of all users do you think engage in the follow behaviors?

- ○ Connecting to an unknown, potentially unsecured public Wi-Fi, and then engaging in sensitive data exchanges, such as transmitting credit card or password details through this connection.
- ○ Reusing same or similar passwords for different accounts
- ○ Downloading illegal or unofficial software/applications and media (e.g., videos, music, and games)
- ○ Ignoring or delaying updating software/applications
- ○ Sharing sensitive personal information online (e.g., location-based information, real-time activities, and pictures of yourself/others) to strangers on social media
- ○ Sharing accounts with family, friends or others

<Q05–Q10: SA-6 questions. A series of SA-6 questions were asked on a 5-point Likert scale: 'strongly disagree,' 'somewhat disagree,' 'neither agree nor disagree,' 'somewhat agree,' and 'strongly agree.'>

Q05. I seek out opportunities to learn about security measures that are relevant to me.

Q06. I am extremely motivated to take all the steps needed to keep my online data and accounts safe.

Q07. Generally, I diligently follow a routine about security practices.

Q08. I often am interested in articles about security threats.

Q09. I always pay attention to experts' advice about the steps I need to take to keep my online data and accounts safe.

Q10. I am extremely knowledgeable about all the steps needed to keep my online data and accounts safe.

Q11. Where do you normally receive your information regarding digital technologies? Please select all that apply.

- ☐ Friends
- ☐ Family
- ☐ School / Teacher
- ☐ Workplace/Colleague
- ☐ News sites / Blogs
- ☐ Service provider / Salesperson
- ☐ Influencers
- ☐ Other (Please Specify)
- ☐ None of the above

Q12. Please select 'Influencers' for this question.

- ☐ Friends
- ☐ Family
- ☐ Colleague
- ☐ Teacher/Mentor
- ☐ Stranger/Online posts
- ☐ Influencers
- ☐ Media (e.g., news websites and TV programs)
- ☐ Service provider/Salesperson
- ☐ Other (Please Specify)
- ☐ I don't remember

Q13. What gender do you identify as?

- ○ Male
- ○ Female
- ○ Non-binary / third gender
- ○ Other (Please Specify)
- ○ Prefer not to say

Q14. How old are you?

Q15. Please select the option which best describes your education level.

- ○ High School or Equivalent
- ○ College diploma
- ○ Undergraduate degree
- ○ Post-graduate education (Masters, Doctorate, Medical/Law School)
- ○ Prefer not to say
- ○ Other (Please Specify)

Q16. Do you consider yourself knowledgeable in Information Technologies or related fields?

- ○ Strongly disagree
- ○ Somewhat disagree
- ○ Neither agree nor disagree
- ○ Somewhat agree
- ○ Strongly agree

Q17. What nationality do you most identify with?

Q18. What country do you currently reside in?

Q19. Please indicate if you have honestly participated in this survey. You will not be penalized/rejected for indicating 'No' but your data may not be included in the final analysis.

- ○ Yes
- ○ No

B Detailed Results

Table 9: Person who advised participants to engage in risky behaviors.

Relationship	%
Family	56.7%
Friend	50.0%
Stranger/Online posts	25.0%
Colleague	20.0%
Media (e.g., news site and TV programs)	10.0%
Influencer	10.0%
Service provider/Salesperson	3.3%
Teacher/Mentor	1.7%
Other	1.7%

Table 9 indicates from whom the participants were advised to engage in risky behaviors. Participants were primarily influenced by family, friends, online strangers, and colleagues, similar to Table 3 (i.e., observation of others).

Beyond Fear and Frustration - Towards a Holistic Understanding of Emotions in Cybersecurity

Alexandra von Preuschen
Justus-Liebig-University Gießen

Monika C. Schuhmacher
Justus-Liebig-University Gießen

Verena Zimmermann
ETH Zurich

Abstract

Employees play a pivotal role for organizational cybersecurity, making understanding the human factor in the context of cybersecurity a critical necessity. While much is known about cognitive factors, less is known about the role of emotions. Through a qualitative survey (N = 112) and in-depth interviews (N = 26), we holistically investigate the causes, types and consequences of emotions in the context of cybersecurity. We demonstrate the existence of diverse, even conflicting emotions at the same time and classify these emotions based on the circumplex model of affect. Furthermore, our findings reveal that essential causes for cybersecurity-related emotions include individual, interpersonal and organizational factors. We also discover various cybersecurity-relevant consequences across behavioral, cognitive and social dimensions. Based on our findings, we provide a framework that unravels the complexity, impact and spill-over effects of cybersecurity-related emotions. Finally, we provide recommendations for promoting secure behavior with a human-centered lens, mitigating negative tendencies, and safeguarding users from unfavorable spill-over effects.

1 Introduction

For decades, the human factor has been considered the weakest link in organizational cybersecurity, often dismissed as lazy or demotivated [23, 84]. This perception has frequently resulted in cumbersome security processes or the use of fear appeals to enforce security guidelines [7, 35, 90]. These everyday experiences with cybersecurity likely cause a spectrum of emotions associated with the term which, in turn, might impact cybersecurity behavior.

As our acknowledgment of humans as integral components of organizational socio-technical systems deepens, there is an increasing importance in understanding human interaction with cybersecurity [17, 54, 76, 83, 90]. In organizational contexts, understanding employee contributions to cybersecurity and the related role of emotions is crucial to protect both companies and the well-being of the employees themselves. Insights from studies exploring the broader impact of emotions in areas such as decision-making, memory and learning, attitude change, or workplace dynamics in general [4, 50, 51, 69, 70], demonstrate the significant and far-reaching impact of emotions in shaping individual actions and cognition towards an object [41, 49].

In the field of cybersecurity, preliminary research also indicates a significant impact of emotions on preventive measures, compliance, and behavioral intentions [6, 16, 22, 35]. Notably, a study by Burns et al. [22] demonstrates that anxiety prompts psychological distancing from cybersecurity, resulting in decreased preventive security measures, while interest leads to the expansion of psychological capabilities, thereby increasing the manifestation of preventive security behavior. Consequently, acknowledging and comprehending cybersecurity experiences and their resulting emotions as well as their consequences is a crucial necessity.

Despite these insights, existing studies related to emotions in cybersecurity exhibit heterogeneity, sometimes contradictory results, mainly focus on negative emotions, particularly fear, and often neglect the complexity of emotions occurring [88]. Consequently, a notable gap persists in the comprehensive understanding of emotions in the context of cybersecurity, including their causes and consequences.

Against this background, this research seeks to close the existing gap by exploring the role of emotions in the context of organizational cybersecurity. To that end, we captured first-hand emotional experiences of employees including experts' as well as employee perspective through a qualitative survey (n = 112) and in-depth interviews (n = 26) that can account for the complexity of emotions. For a holistic understanding, we applied a multi-method approach in the interviews exploring emotions related to cybersecurity in general and specific cybersecurity areas in a multi-faceted way: a) verbally, b) through a non-verbal Product Emotion Measurement Instrument (PrEmo [33, 34]), c) through emotion-related word lists, and d) ratings of emotion intensity. Further, to navigate the complexity of emotions, we applied the circumplex model of

affect [73]. Additionally, emotion causes and consequences were explored. As we know little on how emotions are caused, which emotions occur and what consequences result from them in the context of cybersecurity behavior, we adopt an exploratory and phenomenological qualitative approach. This methodological choice allowed for addressing the complexity of the research topic, while opening the problem space to empathize with employees and to identify emerging patterns [67]. Overall, we investigate three research questions (RQs):

RQ1: Which emotions do employees perceive towards organizational cybersecurity?

RQ2: What causes emotions in the context of organizational cybersecurity?

RQ3: What are the consequences of emotions in organizational cybersecurity?

Our findings show that emotions are caused by four essential themes: individual perceptions, cybersecurity perceptions, interpersonal factors, and organizational factors. Further, we identified multiple emotions towards cybersecurity, extending prior literature. Participants not only but predominantly expressed negatively valenced emotions and overall low-arousal emotions (e.g., 'fearful') were more common than high-arousal ones (e.g., 'interested'). Finally, we find various impacts of cybersecurity-related emotions on individual's cybersecurity perceptions and behaviors, that even extend to other areas of life.

The contribution of our research is three-fold: 1.) We offer a holistic and in-depth exploration of the role of emotions in cybersecurity by employing a multi-modal approach; 2.) Our study develops a theoretical model in the analysis of causes, consequences, and emotions classifying a wide spectrum of cybersecurity-related emotions; and 3.) We provide recommendations for practitioners to enhance favorable consequences, mitigate unfavorable ones among employees, and maintain employees' mental health.

2 Related Work

The following section introduces the concept of emotions and the current state of emotion research within cybersecurity.

2.1 The concept of emotions

Despite the common misconception that emotions are subjective and unpredictable, research demonstrates that affective reactions are often more similar across individuals than cognitive evaluations [72]. Nevertheless, the oversimplification of the concept of 'affect', 'mood' and 'emotion' is a common challenge, often resulting in the terms being used interchangeably [15, 38, 82] with 'affect' often serving as an umbrella term for 'mood' and 'emotion' [28, 73]. 'Mood' is unrelated to specific objects, yet, can result from an emotion when maintained over a longer time [41, 49]. In contrast, emotions, such as happiness or anger, describe an individual's mental state

based on a reaction to a person, event, or object, preparing for action and serving a social function [41]. Feelings, unlike emotions, are purely mental and involve sensations like touch, which are compared to past experiences [60, 86]. Emotions, in turn, express these feelings and are eventually placed in a social context [37, 86]. According to the theory of constructed emotions, emotions are not pre-wired, universal responses to stimuli. Instead, they are actively constructed by the brain based on past experiences, contextual cues, and sensory input [11]. While some theories view emotions as responses to triggers or cognitive evaluations, leading to universal behavioral strategies (e.g., fear triggering a specific facial expression followed by flight behavior [38, 42]), the theory of constructed emotions emphasizes the diversity in emotional experiences and their subsequent actions [12]. Here, emotions describe the result of a process that categorizes sensations by drawing on past experiences and creating situational conceptualizations that best fit the current situation and bodily needs to ultimately guide action [10, 13]. Thus, there is the option to induce emotion consciously, for example by the use of fear appeals to modify behavioral tendencies [58].

Various frameworks for classifying emotions exist such as the circumplex model of affect that offers a structured classification of emotions based on two key dimensions: The vertical axis 'valence' refers to a stimulus's pleasantness ranging from negative to positive; the horizontal axis 'arousal' describes a stimulus' intensity, or the degree of activation of the organism, i.e., mobilization of energy. [56, 73, 81, 82]. For example, the emotion 'sadness' is characterized by a negative valence with a moderate level of arousal [73]. Overall, while emotion theories differ in their processes and terminology, they share a common thread in describing emotions caused by the interpretation of previous experiences and bodily states to prepare for action [8, 57].

Following, we define emotions as mental states resulting from the anticipation of emotional responses that are based on previous emotional experience, the current interpretation of bodily states, perceptions, and environmental cues (e.g., the experience of incidents in the past and cues that are similar in the current state; termed "causes"). They serve the purpose of guiding an individual's action and aiding in prioritizing and organizing behaviors to adapt to environmental demands (e.g., prevention of cognitive overload or maintaining social acceptance; termed "consequences"). Therefore, when analysing emotions in cybersecurity, it is essential to consider their causes and consequences at the same time.

2.2 Emotions in Cybersecurity

Emotions. Most emotion research in the field of cybersecurity derives specific emotions from related fields such as IT usage [22]. Here, studies predominately examine the effect of fear, sadness, or anxiety, mostly using quantitative methods to capture emotions [1, 22, 25, 59]. Furthermore, some research

faces challenges in precisely defining emotion terms, leading to difficulties in adequately capturing emotions [88].

Causes. Current research on the causes of cybersecurity-related emotions is fragmented. Identified causes include cybersecurity incidents [6,21], employer error management [77], the relationship of users and professionals [63], security notifications [29] and persuasive strategies in cybersecurity awareness and education [35,45,89].

Consequences. Initial studies identify emotions and affect as central drivers of behavior within cybersecurity. Studies, for instance, indicate that positive emotions display mixed behavioral tendencies [16,22], with some emotions, notably interest, playing a constructive role in promoting preventive cybersecurity behavior. Other positive emotions such as happiness, as a state of contentment with the current situation, can result in decreased precaution-taking [22]. Negative emotions, in contrast, tend to lead to less favorable behavioral tendencies, often manifesting in avoidance strategies [1, 16, 22]. Yet, results prove to be heterogeneous. While fear has been identified as a deterrent to precaution taking, anxiety may promote favorable cybersecurity behavior such as information-seeking behavior, contributing to an overall sense of precaution [6,22,25]. Similarly, research shows that 'shame' prompts negative actions while 'guilt' can foster self-acceptance and learning [77].

These contradictory results are particularly highlighted when considering induced emotions. Studies show that positive emotional appeals are more effective in promoting stronger password practices compared to negative appeals [45]. Inducing negative emotions such as with fear appeals demonstrate short-term positive effects on security behavior only if coupled with additional factors such as the strengthening of self-efficacy. Nevertheless, despite the eventual positive short-term impact, fear appeals may evoke negative emotions like fear or sadness towards cybersecurity overall that may result in avoidance, decreased well-being, or fear fatigue in the long-term [35,75,89]. While research on the consequences of emotions beyond cybersecurity behavior is limited, there are studies demonstrating that negative emotions in cybersecurity contribute to phenomena like cybersecurity fatigue and burnout [30,72].

Despite the growing interest in emotions within cybersecurity, existing findings display heterogeneity and limitations in capturing the full spectrum of emotions. Furthermore, a holistic understanding of causes and consequences including emotional spill-over effects as a result of cybersecurity-related emotions is currently lacking. Our study addresses this gap by applying a holistic qualitative approach that includes multi-faceted emotion-related measures to unravel the complexity of cybersecurity emotions and their related causes and consequences. Furthermore, we build on the established circumplex model of affect [73] to structure our findings in a meaningful way to inform measures targeted at cybersecurity emotions.

3 Method

The study employed a multi-modal approach, combining semi-structured in-depth interviews and a qualitative survey with overall N=138 participants. This approach allows for qualitatively addressing the complexity of the research topic while exploring emotions with a large number of employees. According to the theory of constructed emotions, verbal reports are essential for assessing the content of subjective emotional experiences as objective measures cannot serve as proxies for emotional experiences [74]. Qualitative surveys complement interviews by mitigating the influence of potential interviewer effects [55]. This strategy aims to overcome the limitations associated with existing research zooming in on a few emotions and the limitations of single methods [74].

3.1 Participants

As we aimed to capture diverse organizational settings, thereby mitigating potential influences of company culture, our recruiting strategy pursued an employee sample of maximum variation including experts' as well as employees' perspectives [68]. We controlled for employee age, cybersecurity background (cybersecurity incident experience, knowledge, attitude, behavior) and organisational background (industry, function, level, security culture). For the interviews, emotional intelligence (EI) was measured to ensure participant's capability to reflect, express and discuss emotions. For details on the variables captured in each study, refer to Appendices B and C. For the recruiting, professionals from different business departments, varying across ranks and industries were approached via participant mailing lists, word-of-mouth, social media (facebook, linkedin, reddit), personal contacts, and snowballing for both the interview and survey. Participants engaged voluntarily and were not financially remunerated for their contributions. Age and work experience were collected in categories to ensure participant's privacy (please refer to 3.3 for a detailed description of ethical aspects).

Qualitative Survey. Our qualitative survey involved 112 participants across at least 18 industries, with 32 identifying as female, 78 as male and 2 as non-binary, varying in age from 18 to 64, and spanning diverse company sizes from 1 to over 1000 employees (referred to as "S_P01-112"). Table 4 shows the comprehensive sample and screening information.

Interview study. The interview study sample consisted of 26 participants of whom 11 identified as female and 15 as male, varying in age from 18 to 64. The sample covered 12 industries with a work experience ranging from 1 to 40 years (referred to as "I_P01-26"). On a seven-point scale, participants rated their IT-expertise with $M = 4.45$ ($SD = 1.30$) and cybersecurity-expertise with $M = 3.77$ ($SD = 1.34$). Data collection was stopped as soon as theoretical saturation was reached [44]. For comprehensive sample information including the sample screening see Table 2.

3.2 Study procedure

Qualitative Survey. For screening of the sample, participants' cybersecurity attitude (SA-6; [39]) and behavioral intention was measured (SeBIS; [36]). Then, participants provided consent and reflected on their (1) emotions towards cybersecurity, (2) thoughts on cybersecurity, (3) cyberattack incident experiences, and provided (4) demographic data. Please refer to Appendix C for detailed information on the survey.

Interviews. Due to the emotion-related nature of this research, physical and psychological safety was considered by informing participants in advance that they were to participate virtually from a safe location and by ensuring that all data was kept confidential to create a comfortable atmosphere that would increase trust and thus to increase the willingness to share information [61]. During the interviews, we used miro - a digital whiteboard - to capture relevant information onto a prepared template, so that the interviewer and interviewee could refer to it throughout the interview. The interview length ranged from 0:24 to 1:27 hours ($M = 0:52$). Before the interview each participant was informed about the objectives, procedures, and data processing of the study and provided informed consent (see Ethical Considerations). Furthermore, for the screening before the interview, they filled out a survey, in which their demographic data was collected first. Then, the survey asked for emotional intelligence using the self-rated emotional intelligence scale [87]. Regarding cybersecurity, knowledge, attitudes and behavior were assessed using an excerpt from the Human Aspects of Information Security Questionnaire (areas from HAIS-Q: password management, email use, internet use) [66] and the climate about cybersecurity was recorded using the Information Security Climate Index (ISCI) [52].

The interview guide was divided into four focus areas detailed in Appendix B:

1) *Emotions towards cybersecurity.* The first focus area aimed to examine emotions towards the general term 'cybersecurity' and its specific areas. Participants were first familiarized with the subject and with the verbalization of emotions by reflecting intuitively on their emotions towards cybersecurity and the relevance of the term 'cybersecurity' in their everyday work. All mentioned emotions were visualized in an emotion-overview in miro. Then, a definition of 'cybersecurity' was introduced to establish a common understanding.

1.a) *General term of cybersecurity.* For a common understanding of the previously described emotions, the participants were presented the non-verbal Product Emotion Measurement Instrument (PrEmo), depicting 14 (7 positive, 7 negative) emotions as cartoons in its second version, to enable participants to reflect thoroughly on their emotions towards cybersecurity [33, 34]. When using the PrEmo, interviewees were instructed to use the tool to help them identify their emotions towards 'cybersecurity' by the use of non-verbal depictions. Thereafter, participants were asked to reflect on the meaning

and perceived intensity on a continuous scale ranging from low to high. To ensure a common understanding, participants were then asked to name the chosen emotion, if possible. After the discussion of the PrEmo, participants were asked to add any further emotions they feel towards cybersecurity, which were not included in the PrEmo. For this, an emotion word list was added to the whiteboard for the supplementation phase after using the PrEmo to facilitate verbalization of emotions that are felt but could not be named ad hoc. For details, see additional digital appendix B (linked in Appendix A). For the creation of the word list, literature was screened for emotions connected with cybersecurity, IT-usage, user experience and basic emotions in general. The number of positively (30) and negatively (30) valenced emotions was balanced and further neutral items (5) were added resulting in a total of 65 emotions. Participants were asked to select three emotions from the prepared word list that best describe their general feelings toward cybersecurity. Both verbal and non-verbal tools were used to help articulate emotions, but participants were not limited to these tools.

1.b) *Specific areas of cybersecurity.* Multiple cybersecurity areas could elicit a variation in emotions (e.g., emotions towards precaution behavior might be different from emotions elicited by a cybersecurity incident) [78] and, thus, influence overall emotions towards cybersecurity. To gain an understanding of emotional experiences influencing the overall emotions towards cybersecurity, we added a section in which participants were asked to reflect on multiple areas within cybersecurity. For this, areas were derived from the user-centered aspects of the NIST framework and visualized in a template on the miro-board [64] However, as capturing emotions retrospectively carries the risk of recall errors and exposes rationalization, a narrative interview section on the main areas was included to encourage participants to rely on their episodic memory [53]. Consequently, participants were guided to reflect in a free narration on their emotional experience within the pre-defined cybersecurity areas, if existent. These emotions were discussed and, if desired, added to the emotion-overview.

2) *Causes and consequences of emotions.* Before delving into the focus area, participants were asked to decide on three emotions that best describe their emotions towards cybersecurity overall. Based on these, we aimed to capture the causes and consequences of participants' emotions towards cybersecurity as a general term. To trigger a change of perspective, a miracle question was additionally used. These questions originate from therapeutic practices, aiming to envision a preferred future rather than holding on to past problems, while encouraging positive changes. Interviewees are asked to imagine how their life would be different if a miracle happened overnight, allowing them to reflect on current shortcomings and needs [32]. Consequences of these three emotions were further asked on both primary (everyday-work) and secondary (cybersecurity) tasks.

3) *Coping.* In the third focus area, we asked participants to reflect on what strategies they use for emotion regulation. We considered intra-individual strategies and strategies of the individual on the part of the company.

4) *Emotions within situational self-efficacy.* The final focus area examined the interdependence of emotions and self-efficacy. Participants were asked to rank their cybersecurity self-efficacy on a scale and describe their reasoning.

3.3 Ethical Considerations

The studies had been reviewed by the independent ethics committee of one author's institution and had been designed to comply with established guidelines for research involving humans [5]. Before both studies, participants were informed about the study's purpose, structure, conditions and data processing, with a clear emphasis on voluntary participation and the right to withdraw without consequences. Participants were informed that participation was voluntary, that they had the right to quit the study at any time without negative consequences and ultimately were asked to read the consent form including the study's data protection policy and give consent. To enhance privacy, we reduced the collection of personal data to a minimum and abstract categories were used. After the interview, each participant received a random identifier for confidentiality, and data were stored on servers complying with national privacy regulations. Overall, the study ensured compliance with national privacy regulations. Considering the potential for participants to share distressing cybersecurity experiences (e.g., feeling ashamed as a result of falling for a phishing email or suffering serious losses due to a cyberattack), interviewers were prepared to handle strong emotions. Interviews could be paused or terminated if necessary, and participants were offered the opportunity to be referred to an appropriate office via the research supervisor for ongoing concerns after the interview.

3.4 Data analysis

All interviews were first transcribed and then analyzed using thematic analysis [19]. As the analysis of complex data benefits from the interaction between coders, multiple interaction and alignment phases were included [65]. First, two coders individually analyzed 20% of the data set that were randomly chosen to derive an initial codebook. Going back and forth several times, a codebook was iteratively developed. A final codebook was formed from discussion and continuous refinement, based on which one researcher coded the complete dataset while aligning with the second coder on the progress multiple times. This approach follows the recommendations for thematic analysis, which advices against multiple independent codings and calculating inter-coder reliability [26]. In the identification of emotions, we also considered terms that are rather cognitive states, feelings, or evaluations (as seen

Figure 1: Eye of cybersecurity-related emotions. See digital appendix A (linked in Appendix A) for a larger color version.

in [31]), and, hence, are not emotions as per definition. Yet, as multiple participants used these terms to describe their emotions, reflecting the subjective and varied nature of emotional experiences in their language and understanding, we integrate terms that are related to emotions (e.g., 'secure'). We omitted participants who expressed emotions related to work-related matters rather than those specifically to cybersecurity.

In a second step, to analyze dependencies, i.e., code configurations of causes, consequences and emotions, we analyzed joint appearances of codes assigned to emotion + consequences or emotion + causes, e.g., exemplary code for cause + emotion: *"Countless passwords. That annoys me. (I_P21)"*.

In a third step, we applied the circumplex model of affect to structure the identified emotions into four classifications (high-low arousal, positive-negative valence) [73]. These classifications were further used for a document-wise reflection on the occurrence of *mixed* emotions across participants.

4 Results

The following sections first introduce the identified emotions with cybersecurity and then describe findings related to the *causes* and the *consequences* of these emotions. Figure 1 provides an illustration of all coded emotions, contextualized in a circumplex model, and their relation to the causes and consequences. Afterward, Figure 2 provides an overview of the underlying framework and the identified emotions, causes, and consequences that align with the section's subheadings.

Following Braun & Clarke's [20] recommendation for reporting results of a thematic analysis, we portray the results of our two studies, provide illustrative quotes and discuss them

directly where applicable. For further quotes, the reader is referred to the codebook in the additional digital Appendix F (number given in brackets (#number)). To avoid the appearance of generalizability and quantification of the answers and to emphasize the depth of the qualitative data, we do not give exact ratios, but instead approximate proportions [20]. Themes and codes that occurred more frequently are provided in descending order.

4.1 Emotions in Cybersecurity

The circumplex model categorizes emotions along the two dimensions: valence (negative - positive) and arousal (low - high) [73]. Overall, participants described more negative than positive emotions with cybersecurity. For positive emotions, participants primarily stated that they feel 'interested', 'secure' (often including feeling self-confident), and 'happy'. For negative emotions, almost all participants stated feeling 'annoyed', whereas almost half of the participants described feeling 'insecure' or 'dependent'. Some participants described emotions that were neither positive nor negative, e.g., being unsure how to feel about the topic. Participants generally described more low-arousal emotions (e.g., 'annoyed', 'uncomfortable' or 'happy'), compared to high-arousal emotions (e.g., 'insecure', 'tense' or 'interested'). For all coded emotions, refer to the gray circle in Figure 1. Almost all participants experienced mixed emotions. For most participants, multiple or all emotion classifications appeared simultaneously (see additional digital appendix E).

4.2 Causes of Emotions in Cybersecurity

4.2.1 Individual Factors: Personal Perceptions

Level of Knowledge and Experience. All participants acknowledged that their level of knowledge and experience influences their emotions toward cybersecurity. The level of knowledge included understanding specific aspects and the general concept of cybersecurity. One person, for example, expressed requiring more knowledge without being able to specify it (#3).

Regarding experience, firstly, emotions were influenced by life experiences, as highlighted by one participant: *"I've been working with computers for about 40 years, and because I've already dealt with many passwords and various things. (I_P11)"*. Secondly, the introduction of new measures or routines triggered emotions (#8), in particular, the experience of receiving suspicious emails (#9). Some noted that emotions tend to become more positive over time with increased experience or routine.

Perceived Level of Protection (active). Many participants reported that their subjective personal engagement and their perceived cybersecurity abilities influenced their emotions (#10). Here, several participants expressed a commitment to

self-defined areas of impact, that do not necessarily align with actual protection levels.

Perceived Lack of Autonomy. Half of the participants expressed limited self-determination in cybersecurity. Specifically, participants felt restricted or coerced by cybersecurity requirements (#11), with some feeling patronized as they lacked the autonomy to decide on the procedure and options of their protection strategy, e.g., time of an update or use of measures such as passwords or biometric authentication: *"I don't have any freedom of choice, I'm just dependent on the arbitrary order to do it that way. (I_P21)"*. Other participants stated that they felt their freedom and rights were generally being curtailed: *"It's a narrative that cybersecurity is an insecure restriction of personal rights. (I_P17)"*.

Internal Conflicts. Most participants expressed internal conflicts involving contradicting attitudes, beliefs, or perceptions. Many described seeing the world as a safe place and a desire to trustfully engage with their environment [27], while simultaneously feeling pressured to adopt a general sense of distrust and experiencing betrayal by individuals they wish to trust. One participant noted: *"I realize that's just the way it is in today's world. You have to be vigilant, you have to be attentive and you have to learn to deal with it. [...] I accept it for myself, even though I don't always like it. (I_P21)"*. Other participants noted a conflict between disinterest and acknowledging cybersecurity's importance or they recognized a discrepancy between their desired and actual engagement in certain behaviors impacting their emotional state.

Perceived Vulnerability. Many participants also reflected on their vulnerability (#15), concerning both, the perceived vulnerability of their company and themselves resulting from behavioral tendencies. Participants often reflected on the extent to which an attack on the company is coincidental to the level of protection (#16).

Anticipated Consequences. The impact of anticipated consequences on participants' emotions varied in terms of the level of abstraction, awareness, and focus. While some reported concrete anticipated consequences, such as business continuity, others depicted rather abstract consequences with far-reaching consequences (#17). Additionally, some participants reflected on the subject of the anticipated consequences being themselves (#19).

Perceived Value of Data. Participants noted that their perception of handled data influences their emotions. In particular, the interviewees reflected on the level of sensitivity of the company's data (#20).

4.2.2 Individual Factors: Cybersecurity Perceptions

Perceived Narrative and Relevance. The participants varied in their perception of cybersecurity's relevance. Many interviewees acknowledged its significance or omnipresence in both their professional and private context (#21). Participants approached cybersecurity from diverse viewpoints, reflecting
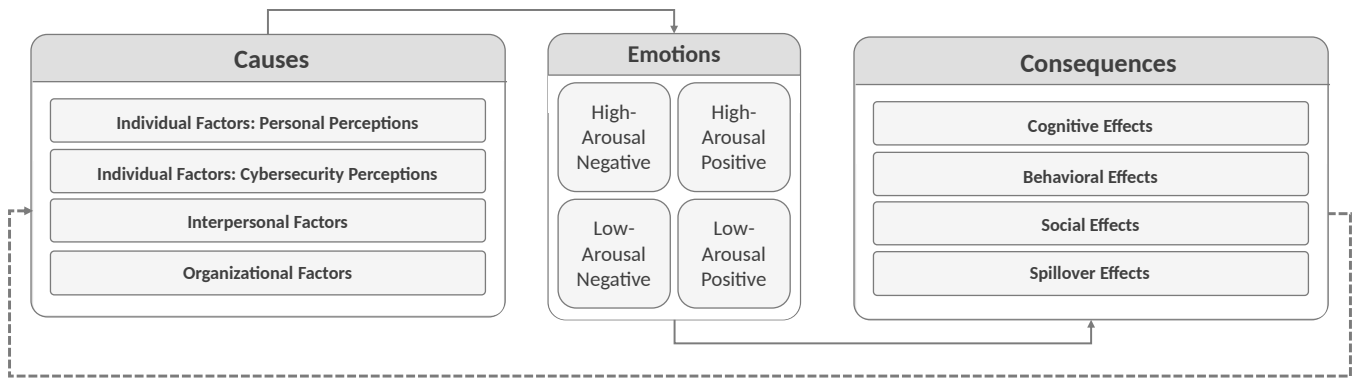
Figure 2: Framework of emotions in cybersecurity

on it both within the context of their company's processes and measures (e.g., password security requirements) and from a broader perspective (e.g., from the point of view of hackers, reporting on attacks, cybersecurity in technical progress): *"On the one hand, I would just be so disinterested when it comes to cybersecurity, but I find that then again I'm interested in how something like that takes place when it comes to things like that, how hackers go about it. (I_P03)"*.

Perceived Resource-Intensiveness and Hindrance. Over half of the participants view cybersecurity as a hindrance or cumbersome to their workflow. They highlighted processes that are perceived as time-consuming or are required at inconvenient times (#24), e.g., password requirements and regular password changes. Furthermore, some participants described a trade-off between security and usability (#23).

Perceived Level of Control. Many participants reflected on their ability to control the possible consequences of cybersecurity attacks, but also on the reliability of security measures which impacts their emotions towards cybersecurity. Some participants delineated aspects where they perceived being able to exert control. Simultaneously, they expressed the limitation of one's influence beyond this defined scope, for example, attacks from unknown parties (#25, 26). The described aspects were often arbitrary and limited to simple basic measures (e.g., locking screens when leaving their workplace). At the same time, some participants described how their own skills are uncontrollable to a certain extent, e.g., influenced by the form of the day, identity or human curiosity: *"I can't do that. [...] I'm not an IT professional. (I_P20)"*. Furthermore, some participants described having only limited influence on preventing an attack among the mass of employees, for example: *"I don't know how many employees we have and yes, my influence is relatively small. (I_P18)"*.

Perceived Level of Necessity. Participants reported different levels of perceived necessity about undertaking cybersecurity measures, e.g. confusion about the purpose of a measure: *"I'm not going to do it. I refused the measure. Out of no understanding of the necessity. (I_P15)")*. Other perceived cybersecurity measures as *"a necessary evil (I_P24)"*. Some

participants described how they feel engaging in cybersecurity is necessary, while others feel that measures are excessive and unnecessary. Some participants generalized this feeling from one measure to the entire concept of cybersecurity.

Perceived Complexity. Some participants outlined that they perceive cybersecurity as such a complex and dull topic that it can only be grasped to a limited extent by everyday users. This perception is similar to parts of the cybersecurity perceptions described by Haney et al. [47]. They also mentioned many technical terms used in the field that are not explained. Some participants also described that no matter how much they learn, there is always more to learn (#31).

Media Reports as Trigger for Cybersecurity Perceptions. Across all individual factors, media reporting was described as the most influential factor for perceptions and, thus, emotions towards cybersecurity. Participants described cybersecurity being portrayed as a negative term with far-reaching consequences for humanity (#32). Some participants outlined that reporting on attacks by related companies in particular triggers emotions.

4.2.3 Interpersonal factors

Self-perception and Perception of Others. Among the most frequently discussed causes for emotions were firstly, the anticipated perception of oneself through colleagues due to cybersecurity behavior or attitudes and secondly, perceptions of colleague's cybersecurity behavior and attitudes. Many participants noted that most colleagues exhibit a low priority for cybersecurity, displaying negative attitudes, substantial knowledge gaps, and insecure behaviors, e.g., *"When I hear the word cybersecurity, the first thing that comes to mind is naivety and stupidity. [...] I also think of ignorance and carelessness. (S_P69)"*. Yet, some participants emphasized sharing the same feeling about cybersecurity with their colleagues. At the same time, many participants expressed concerns about possible negative evaluations such as being seen as paranoid or spoilsports, when exhibiting safe behavior, e.g., *"Maybe I just don't want to describe myself as paranoid.*

(I_P18)"). Furthermore, they worried that their actions may seem inconsistent with their social identity, e.g., *"Sometimes I'm embarrassed about myself, in the sense of what kind of background [IT background] I actually have, whether others know that. How others think about me. [...I] could do better (I_P25)"*. Generational differences in growing up with digital technologies and the subsequent evaluation of one's own and other generations were commonly highlighted (#38, 39).

Level of Social Exchange. While some participants described that the exchange about cybersecurity is an essential part of their work life, the majority expressed a reluctance to talk about cybersecurity. Also, they expressed that others are similarly disinterested in such discussions, e.g., *"Never talked about it, never had the feeling that there was a mood. (I_P20)"*. Yet, many participants noted that they were generally willing to talk about cybersecurity under favorable conditions or when initiated by others.

Perceived Relationship with Experts. More than half of the participants portrayed interpersonal factors shaping the relationship between employees and security experts (or IT department), ultimately influencing emotions in cybersecurity. Participants frequently noted hindered communication characterized by a lack of proactive communication between the two parties, with contacts often initiated in response to negative events (#41). Moreover, they outlined that communication styles including IT-jargon and lengthy explanations, or slow response times create a disconnect with the security department. Other participants perceived being patronized by security experts, akin to the treatment of children: *"Sometimes you really are treated like a small child who just doesn't know how the Internet works yet. (I_P10)"*; *"I think that's more like bullying. (I_P11)"*. Overall, employees expressed feeling undervalued or unappreciated in their efforts and described that their needs are not met. This theme confirms results by Menges et al. [63] showing a dysfunctional relationship between users and experts characterized by particularly negative feelings towards each other, negativity in communication, emotional disengagement and blaming.

4.2.4 Organizational factors

Perceived Level of Protection (passive). While "perceived level of protection (active)" (see section 4.1.1) considers actively taken actions, this theme encompasses actions taken by the company, including technical solutions, availability of policies, and expert support. Many participants articulated the level of trust in the technical solutions provided by their company allowing them to focus on their daily tasks. They also portrayed views on the structural availability of security strategies, reflecting on support options and the overall presence of experts in their infrastructure (#44).

Perception of Design and Frequency of Education. Another subtheme centered around the design and frequency of cybersecurity education, including training materials, commu-

nication, or awareness campaigns. Views on the frequency of educational initiatives varied: Some had a negative perception, especially when content was repetitive, e.g.,: *"I'm annoyed because [...] some things don't need to be told ten times, we know them. (I_P11)"*. This sentiment led to a perceived lack of being taken seriously and a sense of distance from security experts. Some also noted challenges with the complexity of the content and its practical application. Others appreciated frequent training. Notably, some highlighted the importance of their colleagues undergoing training, particularly due to unsafe behavior. Preferences regarding content varied, with some desiring more exciting and fun content, while others questioned the effectiveness of gamification. They expressed a preference of "serious" but well-prepared materials, in particular, due to the seriousness of the topic.

Perceived Security Culture. The perceived importance of security within the company and among colleagues and the priorities by management, shaped participants' perceptions of cybersecurity responsibility at both the team and organizational levels. Some participants felt pressured to adhere to unspoken, potentially insecure guidelines, feeling expectations from colleagues or managers, to conform to such practices, e.g., *"So there are already gray areas being entered to get it done. Then it doesn't matter at that moment. Be it that we break data protection regulations. (I_P22)"*.

Perceived Demands and Requirements. Several participants discussed the burden and practicability of security requirements imposed upon them. Many found security measures and regulations overwhelming and, at times, impractical. While some referred to explicit requirements outlined in policies, others sensed unspoken agreements and expectations that may not align with official security policies (#50).

Error Culture. Many participants referred to the company's error culture, highlighting concerns related to a shaming and blaming culture in the organization, where mistakes are not openly addressed and blamed even if unintentional. Some participants described a secretive organizational culture with no opportunity to learn from others' mistakes: *"But how am I supposed to learn from mistakes if I'm not told about them? (I_P14)"*. Others describe a positive error culture encouraging open discussions about, promoting reporting without fear of reprisal, and prioritizing learning.

4.3 Consequences of Emotions in Cybersecurity

4.3.1 Cognitive Effects

Psychological Distancing and Repression. More than half of the participants showed an unconscious cognitive or emotional separation from the term cybersecurity or consciously suppressed the topic (#52). Distancing oneself from the topic causes disconnection and is associated with a deactivation of positive behavioral tendencies as investigated in the context

of precaution taking [22].

Externalization. Around half of the participants externalized their cybersecurity responsibility, attributing it to their peers, management, security experts, or third-party companies, e.g. for initiating communication and education. On a structural level, many participants demanded or selected technical solutions as a means to abandoning personal responsibility. Some participants described that people with greater expertise should deal with the topic, positioning themselves in a more passive role, e.g., *"I rely on my employer to protect his company. (S_P85)"*.

Distorted Concepts of Cybersecurity and Skills. Some participants narrowed cybersecurity to specific actions, such as avoiding phishing emails, leading to spill-over confidence in broader cybersecurity capabilities. This selective attention contributes to the overestimation of one's overall cybersecurity skills. Furthermore, the impact of incremental improvements is often overestimated (#55).

Level of Self-efficacy. Participants described that their emotions influenced their level of self-efficacy. Nonetheless, a direct connection to emotions was not explicitly articulated (#56). Overall, self-efficacy is known to be highly influenced by emotions [9].

Positive Outcome Expectations. A few participants tended towards convincing themselves of a positive overall situation, and that nothing would happen to them or their company. However, no measures are being taken to ensure that this positive scenario actually occurs. Some showed a tendency to believe that they in comparison to others would be less susceptible to future cyberattacks (e.g. optimism bias, [79, 85]), e.g., *"You know it's somehow not ideal and I hope that nothing will go wrong anyway. (I_P18)"*. This stance is similar to wishful thinking, a belief that is rather based on an individual's desire than actual evidence or rational analysis [14], or optimism bias, a bias underestimating the likelihood of experiencing negative events [18]. Both of which are known to be highly influenced by emotions and investigated in the context of cybersecurity [24, 48]. Yet, optimism bias is known to be independent of cybersecurity education [48].

4.3.2 Behavioral Effects

Level of Attention, Awareness and Caution. Most participants described a shift in the level of their attention between either focusing on a specific area of interest (e.g., potentially harmful emails) or undirected, general attention as a preventative measure without associated measures (#58).

Level and Effectiveness of the Approach to Learning. Half of the participants reflected on the impact of emotions on their willingness and effectiveness to learn. While some described that they actively seek information, others explicitly stated to not seek information. Furthermore, participants outlined the emotion's effect on the effectiveness of learning or retrieving information when needed (#59). Prior research also demonstrated a major effect of emotions on learning, recall, and the effectiveness of academic learning [69].

Avoidance and Rejection. This theme, in contrast to Psychological Distancing and Repression, involves proactive and conscious measures to evade (aspects of) cybersecurity. Half of the participants described that a range of emotions contributes to their avoidance and rejection of specific cybersecurity measures or overall cybersecurity, eventually resulting in a sense of resignation, e.g. *"[This leads to] me not wanting to deal with the issue. And generally not wanting to have anything to do with it (I_P03)"*.

Knowledge-Behavior Gap. Approximately half of the participants admit to not consistently following cybersecurity guidelines, despite being aware of their importance. Some name potential solutions, yet, hesitate to adopt them, e.g., *"I know what these passwords should look like. [...] I usually use a password that I can remember well.[...] Not the super secure ones, I'll admit that. (I_P12)"*.

Security-conscious Behavior. Participants described how cybersecurity had become part of their routine, expressing specific behavioral tendencies or reporting anomalies (#63).

(Concealed) Insecure Behavior. Some participants described engaging in practices that are conducted outside the official security policies of their organization or find workarounds to the company's requirements, yet, are seemingly security-conscious (e.g., having a strong password, but written down: *"I have my file where I write it down. [...] I don't have them all saved in my head (I_P21)"*). In contrast, other participants openly pursue insecure behavior. These behaviors are in line with tendencies revealed by Beris at al. [16] as a consequence of affect.

4.3.3 Social Effects

Level of Social Support Seeking. Participants varied in their active pursuit or desire of social support. This phenomenon includes seeking emotional support, e.g., venting, in line with [59]. An example was: *"When I'm really angry, I can also vent my anger in our office. Then I always get approval. If you're angry, you're not angry alone. [...] And then I'm doing quite well (I_P16)"*. Outward emotion-focused coping, i.e. venting, is associated with increased levels of desirable security behaviors [59]. Some participants, exhibiting low levels of seeking social support, expressed concerns about being perceived negatively, e.g., as paranoid, by others: *"Nowadays, when I say IT or cybersecurity, it has a negative connotation. And that's why I try to avoid the term (I_P14)"*.

Level of Communality. The level of communality is the degree of active support among colleagues. Some participants described actively approaching colleagues to share their knowledge and to work together on cybersecurity (#67). Others described deliberately hiding their knowledge, which has been observed for the interaction between users with high and low cybersecurity expertise [43].

4.3.4 Spillover Effects

Emotional Exhaustion. More than half of the participants described that their emotions towards cybersecurity had far-reaching effects, manifesting in feelings of fear, avoidance of certain topics or tasks, and an overarching sense of burden. One participant noted: *"Sooner or later, it ensures that if this emotion were permanent it would turn into a kind of aversion and therefore the measures are not implemented. (I_P15)"*. Fear, particularly, is seen as a constraint in personal growth (#69). Negative emotions led to prioritization of enjoyable activities over tasks evoking negative emotions. One participant stated: *"Life [without cybersecurity] would be easier, there would be less stress and certainly less burnout at work. (I_P14)"*. A few participants described negative feelings towards their employer: *"Of course, I'm also angry at my employer for constantly making life difficult for me. (I_P13)"*. Dupuis et al. [35] propose that the evocation of negative emotions can generally have negative effects on well-being or job satisfaction. Our results support and extend these findings by showing effects on far-reaching areas of life and that negative experiences (inclusive cybersecurity) are actively avoided.

Reduced Productivity. Participants highlighted that their emotions towards cybersecurity had an impact on their daily productivity, affecting primary work tasks or adopting new technologies. They felt frustrated and annoyed with the constant need to be vigilant and check for phishing emails, at times, leading to ignoring or directly deleting potentially important mails, e.g., *"If I'm not expecting an email, then I don't pay attention to the emails. [...] And if someone really has something important, they can either send me another email or call me. (I_P12)"*.

Need for Recovery. Some participants articulated a need for a timeout as a consequence of negative emotions caused by cybersecurity (#74). Beyond discontinuing their working task, they suggested various methods for recovery, such as disconnecting from technology, going for walks in nature, and engaging in hobbies or activities that provide relaxation and distraction. Despite the short-term impact, some participants noted that emotions arising from colleagues' non-favorable cybersecurity behavior significantly influenced the decision to changing workplaces.

4.4 Contextualization of Findings: The Circumplex Model of Cybersecurity Emotions

Using the circumplex model of emotions, the following sections bring together identified emotions related to their causes and consequences as illustrated in Figure 1.

4.4.1 Identified Cybersecurity Emotions

Causes of cybersecurity-related emotions are displayed as the inner circle and consequences are visualized on the outer circle within the eye of cybersecurity-related emotions in Figure

1. To illustrate the relationships between emotions and their consequences, paths are depicted in Figure 1 while paths for causes-emotions were excluded for better legibility. In the interest of clarity, pathways for causes-emotions were omitted. Please refer to Table 1 for detailed occurrence patterns of the observed interplay of causes-emotions-consequences. For instance, for a low-arousal negative emotion: a *low level knowledge*, *high anticipated consequences* and *negative self-perception or perception of others* resulted in feeling *fearful* and, thus, *psychological distancing* and *(concealed) insecure behavior* or for a an exemplary path for a low-arousal positive emotion: a *high level of perceived protection (active)*, a *high level of perceived control*, a *high level of perceived protection (passive)* and the perception of the organizational *security culture* leads to *happiness* and consequently, in line with Burns et al. [22] *avoidance and rejection* behaviors.

As expected based on the circumplex model of affect, low-arousal emotions were associated with states of low or no action including psychological distancing, avoidance and rejection, and a knowledge-behavior gap. Similarly, low-arousal but positive emotions were linked to psychological distancing, a knowledge-behavior gap, or externalization. Conversely, high-arousal emotions led to a higher activation, particularly increased levels of communality, and higher effectiveness of the approach to learning (see Figure 1). Yet, both high-arousal classifications risk an increased level of (concealed) insecure behavior (particularly for insecurity, fear, and interest).

In contrast to previous results [22], 'interest' was associated with positive and negative behavioral tendencies as well as consequences actually connected to low-arousal emotions (e.g., a decreased level and effectiveness of the approach to learning) and feeling 'secure' (often including feeling self-confident) which resulted in misconceptions or (concealed) insecure behavior. The unfavorable effect of 'interest' can be partially explained by the forced-compliance paradigm that predicts that individuals required to comply with a task perceived as boring experience cognitive dissonance. Thus, as humans strive for balance, they need to balance out the dissonance either by discontinuing or reassessing the perception of the task [40]. Discontinuing is no attractive option as there is a risk of maintaining one's self-image and perception by others Instead, re-evaluating the task helps maintain self-preservation.

Unlike Beris et al. [16], who identified negative behavioral tendencies for negative affect and mixed behavioral tendencies for positive affect, our results demonstrate both behavioral tendencies for both positive and negative affect. This might be because we considered further behavioral tendencies exceeding compliance. Our results reveal that high-arousal negative emotions have no direct positive effect on behavioral tendencies, but display indirect positive effects such as increased information and social support seeking. Yet, in line with the authors' results, our work shows that employees pursue behaviors that might be seemingly secure. In line with

Renaud et al. [77], we found that shame results in undesirable behavioral tendencies.

Considering spillover consequences, low-arousal emotions with a negative valence resulted in overall reduced productivity and emotional exhaustion. 'Interest' was the only positive emotion that was linked to reduced productivity. Please refer to Figure 1 for an illustration of the interconnections between emotions and consequences.

4.4.2 Mixed Emotions

Despite varying backgrounds, including a variation in knowledge or industry, participants display mixed behavioral and cognitive tendencies of favorable and unfavorable nature. Thus, multiple behavioral tendencies and occasionally contradicting cognitions are present simultaneously stemming from emotional dissonance. For example, participants feel interested in cybersecurity and would like to learn more about it, still, they are afraid of being judged by their colleagues and avoid the topic overall. Another illustrative example: Some participants are knowledgeable, feel secure and would like to engage in secure behavior, yet, feel patronized by security education and consciously act against guidelines. For an details on the document-wise assignment of codes, see digital Appendix E.

5 Discussion

5.1 Summary of Key Results

Overall, our findings shed light on the role of emotions in cybersecurity by highlighting causes, types and consequences of emotions. Delving into the causes of cybersecurity, our study expands upon prior research [45, 63, 77] by categorizing examined factors in four themes: individual personal perceptions, individual cybersecurity perceptions, interpersonal, and organization-wide factors. While existing literature predominantly focuses on negative emotions such as fear, sadness [1, 6, 89], often derived from related areas such as IT usage [22], our exploratory approach presents a comprehensive perspective on the emotions towards cybersecurity. Indeed, feelings of fear and insecurity were highly prevalent, yet, only a small share of the experienced emotions towards cybersecurity overall. While previous research often considered the experience of one single uniform emotion [16, 22, 25], our research reveals the simultaneous occurrence of multiple contradicting emotions in most individuals. This also supports the theory of constructed emotions, explaining the diverse and complex emotions reported, influenced by personal, social, and organizational factors in cybersecurity. While previous research primarily considered behavioral tendencies including precaution behavior, compliance, and emotional coping behavior [16, 22, 25, 59], our results confirm and extend them by revealing a complex interplay of multiple behavioral, cognitive,

and social consequences simultaneously. Furthermore, we show that emotions towards cybersecurity spill-over to other areas of life: some individuals feel emotionally exhausted, impeded in their productivity, or feel a need for distancing from their work in general.

5.2 Recommendations for Cybersecurity Practitioners

Overall, our findings indicate that practitioners should aim for *first* addressing emotions while reducing emotional dissonance (e.g. through the establishment of an emotion-oriented mindset). *Second*, high-arousal emotions and subsequent causes should be enhanced while considering the risk of undesirable activation i.e. *(concealed) insecure behavior* and low-arousal emotions and their subsequent causes should be diminished. We advise for a holistic strategy as emotions caused by one area can impact the overall approach to cybersecurity. This approach seeks to integrate the humans with all their complexities, into the socio-technical framework of organizational cybersecurity. Additionally, it aims to protect individuals from potential negative consequences thereby enhancing their ability to focus on their primary work task. Key components of the advised strategy are the following:

5.2.1 Establishment of an Emotion-oriented mindset

Cultivate empathy. The lack of security behavior or behavior change in general is mostly determined by the perception of emotional ambivalence [80]. Practitioners should recognize the role of emotions and establish channels for emotional support, where employees can share their emotions (anonymously), *seek social support*, foster a positive *sense of cybersecurity culture* and, thus, prevent *emotional exhaustion*. Additionally, cultivating empathy towards experts enhances the *relationship with experts*. We advise to share real-life cybersecurity stories and case studies within the organization to improve *cybersecurity perceptions* and the *expert-user relationship*. As storytelling was already shown to have positive effects on cybersecurity education [71], it might also be leveraged for cultivating empathy.

Set the stage. To mitigate internal conflicts, we recommend creating a culture of psychological safety where employees should feel empowered to ask for expectations and question tasks perceived as insecure. Acceptance of varying interest levels in cybersecurity is crucial, and enforcement strategies should be avoided to prevent suboptimal results. Instead, cybersecurity should be presented in relatable terms, portraying realistic consequences and clearly defining *areas of control*. Recognizing that some employees may perceive their impact as minimal, especially in light of colleagues' insecure behavior, it is crucial to make employees aware that everyone plays a valuable role in the company's security strategy [90].

Foster emotional reflection. While enhancing positive emotions can help overcome negative emotions, there's a potential drawback: the introduction of positive low-arousal emotions associated with undesirable behavioral tendencies. To ensure mental health and emotional resilience, it is crucial to promote emotional reflection to maintain a balanced and healthy emotional state within the cybersecurity context.

5.2.2 Enhancement of high-arousal Emotions and Diminution low-arousal Emotions

Here, we outline exemplary strategies for enhancing high-arousal and mitigating low-arousal emotions. Further strategies can be derived from Figure 1 by examining and modifying causes of low-arousal or high-arousal emotions. For instance, low levels of perceived control were identified as a cause for negative low-arousal emotions and subsequent negative consequences. Providing users with a moderate sense of control through **clear communication**, such as imparting hands-on strategies like emphasizing the importance of password length to prevent brute-force hacking, can convey a sense of control. Further, fostering an environment of transparency, it is crucial to articulate cybersecurity goals, i.e., the *area of control*, and the *necessity* of measures clearly. Employees should feel able to influence security measures such as by giving the possibility to update a software at one of two time-slots. Involving user representatives in decision-making processes enhances a sense of *autonomy* among employees. Yet, attention must be paid to strategy implementation, as high levels of perceived control can result in feelings of positive low-arousal emotions and undesired consequences.

The *level of knowledge* and expertise is a major cause of high-arousal emotions, while also posing the risk of impacting low-arousal emotions. Therefore, we advise carefully fostering high-arousal emotions and mitigating low-arousal emotions, such as through the implementation of **individualized cybersecurity education**. While some employees struggle with IT-jargon, others feel bored or coerced by repetitive or basic training (*perception of design an frequency of education*). Thus, we recommend assessing the learner's knowledge level and offering training tailored to their needs as recently proposed, e.g. by [2, 3]. Furthermore, employees prefer material that is coherent with their emotional tone and perceptions. Thus, not all employees enjoy fun or gamified training. A survey by McLauglin [62] indicates that especially leader boards decrease learning desire. Negative low-arousal emotions often stem from perceived *expertise levels*. To counteract this, we recommend developing educational material grounded in real-world scenarios. However, caution is advised as high levels of perceived expertise or the *perceived level of protection (active)* pose a risk of feeling too secure and, thus, *distorted concepts of cybersecurity*. We recommend fostering regular reflections on skills but also actually implemented measures. However, reflecting on low levels of security behavior might result in a

cognitive dissonance for those with positive emotions. Hence, employees may not be blamed [77] but should be encouraged to view security behaviors as an ongoing improvement process rather than expecting instant changes. This approach mitigates the risk of cognitive dissonance resulting from the misalignment of emotions and implemented behavior. Further, employees with high knowledge or expertise levels can be impeded from openly discussing and engaging with cybersecurity due to concerns about negative perceptions from others (similar as in [43]). To address this challenge, we recommend empowering these employees by designating them as **ambassadors** and providing support to them as Gutfleisch et al. [46] illustrated that mere appointment of "security champions" without management and IT support is insufficient.

Considering the examined spill-over effects we conclude that **scaring won't do in long-term.** Despite the potential positive short-term effect of fear appeals as seen in prior research [35], scaring employees into compliance may result in fear, negative low-arousal emotions, negative effects on security behavior, the interpersonal and organizational environment and cybersecurity-related perceptions [35]. Thus, fear appeals might motivate short-term secure decisions, however, ultimately result in psychological distancing or even emotional exhaustion. To mitigate these risks, we recommend prioritizing emotional reflection over fear-based approaches.

5.3 Limitations and Future Work

While our study provides valuable insights into the interplay of emotions and cybersecurity, some limitations need to be considered. *First*, our study examined a wide range of emotions in cybersecurity but did not extensively analyze complex dependencies, such as the interplay of multiple causes or consequences of specific emotional constellations.

Second, the exploratory qualitative nature of our study limited the quantification of results. Future research could delve deeper into specific cybersecurity areas, examining emotions and their (co-)dependencies quantitatively. Adopting a mixed methods approach would benefit capturing the complex dynamics around cybersecurity emotions. *Third*, our research took a retrospective view of cybersecurity emotions, potentially overlooking temporal changes. Future research could explore how emotions evolve, e.g., in response to incidents, and their long-term impact on cybersecurity attitudes or behaviors. Further, we acknowledge that cybersecurity-related emotions might overlap with general workplace issues despite aiming for maximum variation in the sample. Our study relied on participants' cybersecurity-focused responses. Thus, future research could explore the interaction between cybersecurity and workplace culture. Future research could also investigate how strategic cybersecurity measures impact these emotions and the related consequences or behaviours, respectively or develop measurement tools that benefit from emotions capturing several causes and consequences simultaneously.

Acknowledgments

We would like to thank Anna-Maria Klein, Miriam Pitzer and Julius Klein for the support in data collection.

Data Availability Statement

Due to the high sensitivity of interview data, we do not make the interview data publicly available. Detailed information on the sample, the interview guide, code book, and exemplary quotes are provided with the article to enhance transparency and replicability. For access to the original interview transcripts, please contact the authors.

References

[1] Hossein Abroshan, Jan Devos, Geert Poels, and Eric Laermans. Covid-19 and phishing: Effects of human emotions, behavior, and demographics on the success of phishing attempts during the pandemic. *Ieee Access*, 9:121916–121929, 2021.

[2] Yusuf Albayram, David Suess, Yassir Yaghzar Elidrissi, Daniel P. Rollins, and Maciej Beclawski. Personalized cybersecurity education: A mobile app proof of concept. In *HCI International 2023 – Late Breaking Posters*, Communications in Computer and Information Science, pages 257–263, Cham, 2024. Springer Nature Switzerland and Imprint Springer.

[3] S Alotaibi, Steven Furnell, and Y He. Towards a framework for the personalization of cybersecurity awareness. In *International Symposium on Human Aspects of Information Security and Assurance*, pages 143–153. Springer, 2023.

[4] Neal M. Ashkanasy and Alana D. Dorris. Emotions in the workplace. *Annual Review of Organizational Psychology and Organizational Behavior*, 4(1):67–90, 2017.

[5] American Psychological Association. Ethical principles of psychologists and code of conduct. https://www.apa.org/ethics/code, 2023. [Online; accessed: 09-February-2024].

[6] Eric Bachura, Rohit Valecha, Rui Chen, and H Raghav Rao. The opm data breach: An investigation of shared emotional reactions on twitter. *MIS Quarterly*, 46(2), 2022.

[7] Maria Bada, Angela M. Sasse, and Jason R. C. Nurse. Cyber security awareness campaigns: Why do they fail to change behaviour? *International Conference on Cyber Security for Sustainable Society*, 2015.

[8] R. P. Bagozzi, M. Gopinath, and P. U. Nyer. The role of emotions in marketing. *Journal of the Academy of Marketing Science*, 27(2):184–206, 1999.

[9] Albert Bandura. Social cognitive theory of personality. *Handbook of personality*, 2:154–96, 1999.

[10] Lisa Feldman Barrett. Solving the emotion paradox: categorization and the experience of emotion. *Personality and social psychology review : an official journal of the Society for Personality and Social Psychology, Inc*, 10(1):20–46, 2006.

[11] Lisa Feldman Barrett. The theory of constructed emotion: an active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, 12(1):1–23, 2017.

[12] Lisa Feldman Barrett and Christiana Westlin. Navigating the science of emotion. In *Emotion measurement*, pages 39–84. Elsevier, 2021.

[13] L. W. Barsalou. Perceptual symbol systems. *Behavioral and Brain Sciences*, 22(4):577–609; discussion 610–60, 1999.

[14] Anthony Bastardi, Eric Luis Uhlmann, and Lee Ross. Wishful thinking: Belief, desire, and the motivated evaluation of scientific evidence. *Psychological science*, 22(6):731, 2011.

[15] Christopher Beedie, Peter Terry, and Andrew Lane. Distinctions between emotion and mood. *Cognition & Emotion*, 19(6):847–878, 2005.

[16] Odette Beris, Adam Beautement, and M Angela Sasse. Employee rule breakers, excuse makers and security champions: mapping the risk perceptions and emotions that drive security behaviors. In *Proceedings of the 2015 New Security Paradigms Workshop*, pages 73–84, 2015.

[17] Scott R Boss, Dennis F Galletta, Paul Benjamin Lowry, Gregory D Moody, and Peter Polak. What do systems users have to fear? using fear appeals to engender threats and fear that motivate protective security behaviors. *MIS quarterly*, 39(4):837–864, 2015.

[18] Anat Bracha and Donald J Brown. Affective decision making: A theory of optimism bias. *Games and Economic Behavior*, 75(1):67–80, 2012.

[19] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

[20] Virginia Braun and Victoria Clarke. *Thematic analysis: A practical guide*. SAGE, Los Angeles, 2022.

[21] Sanja Budimir, Johnny RJ Fontaine, and Etienne B Roesch. Emotional experiences of cybersecurity breach victims. *Cyberpsychology, Behavior, and Social Networking*, 24(9):612–616, 2021.

[22] AJ Burns, Tom L Roberts, Clay Posey, and Paul Benjamin Lowry. The adaptive roles of positive and negative emotions in organizational insiders' security-based precaution taking. *Information Systems Research*, 30(4):1228–1247, 2019.

[23] Perry Carpenter and Kai Roer. *The Security Culture Playbook: An Executive Guide to Reducing Risk and Developing Your Human Defense Layer*. John Wiley & Sons, 2022.

[24] Daniel Qi Chen and Huigang Liang. Wishful thinking and it threat avoidance: An extension to the technology threat avoidance theory. *IEEE Transactions on Engineering Management*, 66(4):552–567, 2019.

[25] Violet Cheung-Blunden, Kiefer Cropper, Aleesa Panis, and Kamilah Davis. Functional divergence of two threat-induced emotions: Fear-based versus anxiety-based cybersecurity preferences. *Emotion*, 19(8):1353, 2019.

[26] Victoria Clarke and Virginia Braun. Successful qualitative research: A practical guide for beginners. *Successful qualitative research*, pages 1–400, 2013.

[27] Jeremy DW Clifton, Joshua D Baker, Crystal L Park, David B Yaden, Alicia BW Clifton, Paolo Terni, Jessica L Miller, Guang Zeng, Salvatore Giorgi, H Andrew Schwartz, et al. Primal world beliefs. *Psychological Assessment*, 31(1):82, 2019.

[28] Gerald L Clore, Norbert Schwarz, and Michael Conway. Affective causes and consequences of social information processing. In *Handbook of social cognition*, pages 323–418. Psychology Press, 2014.

[29] Colin D. Conrad, Jasmine R. Aziz, Jonathon M. Henneberry, and Aaron J. Newman. Do emotions influence safe browsing? toward an electroencephalography marker of affective responses to cybersecurity notifications. *Frontiers in Neuroscience*, 16:922960, 2022.

[30] W Alec Cram, Jeffrey G Proudfoot, and John D'Arcy. When enough is enough: Investigating the antecedents and consequences of information security fatigue. *Information Systems Journal*, 31(4):521–549, 2021.

[31] Cynthia D. Fisher. *What do people feel and how should we measure it?* Bond University - School of Business Discussion Papers, 1997.

[32] Steve De Shazer, Yvonne Dolan, Harry Korman, Terry Trepper, Eric McCollum, and Insoo Kim Berg. *More than miracles: The state of the art of solution-focused brief therapy*. Routledge, 2021.

[33] Pieter Desmet. Measuring emotion: Development and application of an instrument to measure emotional responses to products. *Funology 2*, pages 391–404, 2018.

[34] Pieter Desmet, Peter Wassink, and Yancheng Du. Premo (emotion measurement instrument) card set: Male version, 2019.

[35] Marc Dupuis, Karen Renaud, and Anna Jennings. Fear might motivate secure password choices in the short term, but at what cost? In *Hawaii International Conference on System Sciences*, 2021.

[36] Serge Egelman and Eyal Peer. Scaling the security wall: Developing a security behavior intentions scale (sebis). In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 2873–2882, 2015.

[37] Paul Ekman. Universals and cultural differences in facial expressions of emotion. In *Nebraska symposium on motivation*. University of Nebraska Press, 1971.

[38] Paul Ed Ekman and Richard J Davidson. *The nature of emotion: Fundamental questions.* Oxford University Press, 1994.

[39] Cori Faklaris, Laura A Dabbish, and Jason I Hong. A self-report measure of end-user security attitudes (sa-6). In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, pages 61–77, 2019.

[40] Leon Festinger and James M Carlsmith. Cognitive consequences of forced compliance. *The journal of abnormal and social psychology*, 58(2):203, 1959.

[41] Nico H. Frijda. Moods, emotion episodes, and emotions. In *Handbook of emotions*, pages 381–403. The Guilford Press, New York, NY, US, 1993.

[42] Nico H. Frijda, Peter Kuipers, and Elisabeth ter Schure. Relations among emotion, appraisal, and emotional action readiness. *Journal of Personality and Social Psychology*, 57(2):212–228, 1989.

[43] Nina Gerber and Karola Marky. The nerd factor: The potential of S&P adepts to serve as a social resource in the user's quest for more secure and Privacy-Preserving behavior. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 57–76, Boston, MA, August 2022. USENIX Association.

[44] Greg Guest, Arwen Bunce, and Laura Johnson. How many interviews are enough? *Field Methods*, 18(1):59–82, 2006.

[45] Iwan Gulenko. Improving passwords: Influence of emotions on security behaviour. *Information Management & Computer Security*, 22(2):167–178, 2014.

[46] Marco Gutfleisch, Markus Schöps, Stefan Albert Horstmann, Daniel Wichmann, and M Angela Sasse. Security champions without support: Results from a case study with owasp samm in a large-scale e-commerce enterprise. In *Proceedings of the 2023 European Symposium on Usable Security*, pages 260–276, 2023.

[47] Julie M. Haney and Wayne G. Lutters. "it's Scary. . . It's Confusing. . . It's dull": How cybersecurity advocates overcome negative perceptions of security. In *Fourteenth Symposium on Usable Privacy and Security*, SOUPS 2018, pages 411–425, Baltimore, MD, August 2018. USENIX Association.

[48] Barbara Hewitt and Garry L White. Optimistic bias and exposure affect security incidents on home computer. *Journal of Computer Information Systems*, 62(1):50–60, 2022.

[49] Alice M Isen. *Toward understanding the role of affect in cognition.* Lawrence Erlbaum Associates Publishers, 1984.

[50] Daniel Kahneman. *Thinking, fast and slow.* macmillan, 2011.

[51] Elizabeth A Kensinger and Jaclyn H Ford. Retrieval of emotional events from memory. *Annual review of psychology*, 71:251–272, 2020.

[52] Stacey R Kessler, Shani Pindek, Gary Kleinman, Stephanie A Andel, and Paul E Spector. Information security climate and the assessment of information security risk among healthcare employees. *Health informatics journal*, 26(1):461–473, 2020.

[53] Saouré Kouamé and Feng Liu. Capturing emotions in qualitative strategic organization research. *Strategic Organization*, 19(1):97–112, 2021.

[54] Sara Kraemer and Pascale Carayon. Human errors and violations in computer and information security: The viewpoint of network administrators and security specialists. *Applied ergonomics*, 38(2):143–154, 2007.

[55] Ivar Krumpal. Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & quantity*, 47(4):2025–2047, 2013.

[56] Peter J Lang, Margaret M Bradley, and Bruce N Cuthbert. Emotion, attention, and the startle reflex. *Psychological review*, 97(3):377, 1990.

[57] Richard S. Lazarus. *Emotion and Adaptation.* Oxford University Press, 1991.

[58] Howard Leventhal. Findings and theory in the study of fear communications. *Advances in experimental social psychology*, 5:119–186, 1970.

[59] Huigang Liang, Yajiong Xue, Alain Pinsonneault, and Yu Andy Wu. What users do besides problem-focused coping when facing it security threats: An emotion-focused coping perspective. *MIS quarterly*, 43(2):373–394, 2019.

[60] Catherine A Lutz. *Unnatural emotions: Everyday sentiments on a Micronesian atoll and their challenge to Western theory.* University of Chicago Press, 2011.

[61] Heather McCosker, Alan Barnard, and Rod Gerber. Undertaking sensitive research: Issues and strategies for meeting the safety needs of all participants. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, 2(1), 2001.

[62] Kevin Mclaughlin. *A Quantitative Study of Learner Choice in Cybersecurity Training: Do They Even Want Gamification?* PhD thesis, Colorado Technical University, 2023.

[63] Uta Menges, Jonas Hielscher, Annalina Buckmann, Annette Kluge, M Angela Sasse, and Imogen Verret. Why it security needs therapy. In *European Symposium on Research in Computer Security*, pages 335–356. Springer, 2021.

[64] NIST (National Institute of Standards and Technology). Framework for improving critical infrastructure cybersecurity, 2014.

[65] Anna-Marie Ortloff, Matthias Fassl, Alexander Ponticello, Florin Martius, Anne Mertens, Katharina Krombholz, and Matthew Smith. Different researchers, different results? analyzing the influence of researcher experience and data type during qualitative analysis of an interview and survey study on security advice. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2023.

[66] Kathryn Parsons, Dragana Calic, Malcolm Pattinson, Marcus Butavicius, Agata McCormac, and Tara Zwaans. The human aspects of information security questionnaire (hais-q): two further validation studies. *Computers & Security*, 66:40–51, 2017.

[67] Michael Quinn Patton. Qualitative research and evaluation methods. thousand oaks. *Cal.: Sage Publications*, 4, 2002.

[68] Michael Quinn Patton. Two decades of developments in qualitative inquiry: A personal, experiential perspective. *Qualitative social work*, 1(3):261–283, 2002.

[69] Reinhard Pekrun, Thomas Goetz, Wolfram Titz, and Raymond P Perry. Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational psychologist*, 37(2):91–105, 2002.

[70] Richard E Petty and Pablo Briñol. Emotion and persuasion: Cognitive and meta-cognitive processes impact attitudes. *Cognition and Emotion*, 29(1):1–26, 2015.

[71] Katharina Pfeffer, Alexandra Mai, Edgar Weippl, Emilee Rader, and Katharina Krombholz. Replication: Stories as informal lessons about security. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 1–18, Boston, MA, August 2022. USENIX Association.

[72] Michel Tuan Pham, Joel B Cohen, John W Pracejus, and G David Hughes. Affect monitoring and the primacy of feelings in judgment. *Journal of consumer research*, 28(2):167–188, 2001.

[73] Jonathan Posner, James A Russell, and Bradley S Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, 17(3):715–734, 2005.

[74] Karen S Quigley, Kristen A Lindquist, and Lisa Feldman Barrett. Inducing and measuring emotion and affect: Tips, tricks, and secrets. *Handbook of research methods in social and personality psychology*, 220:252, 2014.

[75] Karen Renaud and Marc Dupuis. Cyber security fear appeals: Unexpectedly complicated. In *Proceedings of the new security paradigms workshop*, pages 42–56, 2019.

[76] Karen Renaud and Stephen Flowerday. Contemplating human-centred security & privacy research: Suggesting future directions. *Journal of Information Security and Applications*, 34:76–81, 2017.

[77] Karen Renaud, Rosalind Searle, and Marc Dupuis. Shame in cyber security: effective behavior modification tool or counterproductive foil? In *New Security Paradigms Workshop*, pages 70–87, 2021.

[78] Karen Renaud, Verena Zimmermann, Tim Schürmann, and Carlos Böhm. Exploring cybersecurity-related emotions and finding that they are challenging to measure. *Humanities and Social Sciences Communications*, 8(1):1–17, 2021.

[79] Hyeun-Suk Rhee, Young Ryu, and Cheong-Tag Kim. I am fine but you are not: Optimistic bias and illusion of control on information security. *ICIS 2005 proceedings*, page 32, 2005.

[80] Naomi B Rothman, Michael G Pratt, Laura Rees, and Timothy J Vogus. Understanding the dual nature of ambivalence: Why and when ambivalence leads to good and bad outcomes. *Academy of Management Annals*, 11(1):33–72, 2017.

[81] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[82] James A Russell and Lisa Feldman Barrett. Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of personality and social psychology*, 76(5):805, 1999.

[83] Martina Angela Sasse, Sacha Brostoff, and Dirk Weirich. Transforming the 'weakest link'—a human/computer interaction approach to usable and effective security. *BT technology journal*, 19(3):122–131, 2001.

[84] Bruce Schneier. *Secrets and lies: digital security in a networked world.* John Wiley & Sons, 2015.

[85] Tali Sharot, Alison M Riccardi, Candace M Raio, and Elizabeth A Phelps. Neural mechanisms mediating optimism bias. *Nature*, 450(7166):102–105, 2007.

[86] Eric Shouse. Feeling, emotion, affect. *M/c journal*, 8(6), 2005.

[87] Matthias Vöhringer, Astrid Schütz, Sarah Gessler, and Michela Schröder-Abé. Sreis-d. *Diagnostica*, 66(3):200–210, 2020.

[88] Alexandra von Preuschen, Verena Zimmermann, and Monika C Schuhmacher. How do you feel about cybersecurity? - a literature review on emotions in cybersecurity. *Proceedings TecPsy 2023*, page 1, 2023.

[89] Xiaochen Angela Zhang and Jonathan Borden. How to communicate cyber-risk? an examination of behavioral recommendations in cybersecurity crises. *Journal of Risk Research*, 23(10):1336–1352, 2020.

[90] Verena Zimmermann and Karen Renaud. Moving from a 'human-as-problem" to a 'human-as-solution" cybersecurity mindset. *International Journal of Human-Computer Studies*, 131:169–187, 2019.

A Appendix: Data Analysis

Further supplementary material including an enlarged color version of the eye of cybersecurity-related emotions, an depiction of the causes (inner circle), analyses on mixed emotions and our codebook is available at: https://www.research-collection.ethz.ch/handle/20.500.11850/669758

Table 1 (rotated). Columns are emotions grouped as:

POSITIVE — *high-arousal:* interested, fascinated, empathetic, action-motivated, involved; *low-arousal:* sympathetic, calm, trustful, hopeful, happy

NEGATIVE — *high-arousal:* secure, astonished, threatened, tense, insecure, dependent, left alone; *low-arousal:* fatigued, indifferent, impotent, overwhelmed, bored, ashamed, dull, dismissive, averse, frustrated, uncomfortable, spiteful, annoyed, disappointed, helpless, powerless, fearful

CAUSES

	fearful	powerless	helpless	disappointed	annoyed	spiteful	uncomfortable	frustrated	averse	dismissive	dull	ashamed	bored	overwhelmed	impotent	indifferent	fatigued	left alone	dependent	insecure	tense	threatened	astonished	secure	happy	hopeful	trustful	calm	sympathetic	involved	action-motivated	empathetic	fascinated	interested
Individual factors: Personal perceptions																																		
Level of Knowledge and Experience	•		•		•						•		•	•					•	•	•	•	•	•	•		•	•	•	•		•		•
Perceived Level of Protection (active)					•								•		•				•			•		•	•	•		•	•					
Perceived Lack of Autonomy			•		•		•													•														
Internal Conflicts																•					•		•											
Perceived Vulnerability	•	•	•															•											•					•
Anticipated Consequences	•																																	
Perceived Value of Data																									•				•	•			•	•
Individual factors: Cybersecurity Perceptions																																		
Perceived Narrative and Relevance					•			•									•			•	•													
Perceived Resource-Intensiveness and Hindrance					•			•			•									•	•								•				•	
Perceived Level of Control			•		•										•				•	•	•			•	•		•		•					
Perceived Level of Necessity					•																•				•				•					
Perceived Complexity		•											•																					
Media Reports as a Trigger for Cybersecurity Perceptions																					•													•
Interpersonal factors																																		
Self-perception and Perception of Others	•				•		•	•								•							•		•	•						•		•
Level of Social Exchange					•	•	•	•										•																
Perceived Relationship with Experts						•											•				•								•					•
Organization-wide factors																																		
Perceived Level of Protection (passive)	•	•			•	•		•					•				•	•	•	•				•	•	•	•		•				•	
Perception of Design and Frequency of Education					•	•												•						•	•				•					
Perceived Security Culture					•	•						•			•				•		•			•	•	•								
Perceived Demands and Requirements		•																		•	•					•								
Error Culture	•						•														•			•		•			•					

CONSEQUENCES

	fearful	powerless	helpless	disappointed	annoyed	spiteful	uncomfortable	frustrated	averse	dismissive	dull	ashamed	bored	overwhelmed	impotent	indifferent	fatigued	left alone	dependent	insecure	tense	threatened	astonished	secure	happy	hopeful	trustful	calm	sympathetic	involved	action-motivated	empathetic	fascinated	interested
Cognitive																																		
Psychological Distancing and Repression	•							•				•			•									•	•	•								
Externalization		•																																
Distorted concepts of cybersecurity and skills																								•		•								
Level of self-efficacy																				•		•	•			•								•
Positive Outcome expectations																	•														•			
Behavioral																																		
Level of Attention, Awareness and Caution			•		•			•					•			•	•																	
Level and effectiveness of the approach to learning					•		•	•					•			•	•																	
Avoidance and Rejection					•		•	•				•								•														
Knowledge-Behavior Gap																								•		•				•				
Security-conscious Behavior	•																	•		•				•		•			•			•		•
(Concealed) insecure Behavior	•						•																											
Social																																		
Level of Social support seeking		•																•																•
Level of communality					•																							•	•			•		•
Spillover																																		
Emotional exhaustion							•										•	•																
Reduced productivity					•													•																
Need for recovery																	•																	

Table 1: Segment-wise analysis of causes and consequences and emotions (uniform emotions).

B   Appendix: Interview

Interview Guideline

Introduction

- Participants were welcomed to the study and introduced to the background of the study

- Participants were reminded of participation conditions, acknowledging potential discomfort. They were encouraged to take time to answer, consider their responses, discontinue if necessary due to strong negative emotions, or seek further support afterward.

    - Spontaneously: When you think of cybersecurity, how does it make you feel?

    - How do you define cybersecurity?

- Interviewer provided a brief definition of the term cybersecurity

Emotions towards cybersecurity

1.a) General term of cybersecurity

- PrEmo was displayed. These questions were repeated until no further illustration showed the felt emotions:

    - Which of these illustrations best shows your feelings about cybersecurity?

    - What does this emotion mean to you?

    - How is this emotion expressed?

    - Can you scale this emotion from low to high on this scale?

    - Can you find a name for this emotion?

- The emotion word list was presented, and participants were instructed to mark feelings they experience, then narrow it down to three terms that best describe their feelings toward cybersecurity.

- Selected emotions were added to the main board. Questions on the understanding of the emotions are repeated if necessary

    - Please try to put yourself in a different position: How do you think your colleagues feel about cybersecurity in the workplace?

1.b Specific areas of cybersecurity

- Specific areas of cybersecurity were explained

    - I would like to ask you to tell me about your experience from your everyday work in relation to these aspects. Share what comes to mind, take as

much time as you need, and please focus on how you felt in these situations. I will not interrupt you for now, but I will be making notes on the side.

Top Emotions

- Participants could add further emotions to the main board if wished

- Three emotions (top emotions) were selected for the further interviewing process

Antecedents

- The following questions were asked:

    - Why do you feel the way you do when you think about cybersecurity (Top 3)?

    - What emotion would you like to feel towards cybersecurity?

    - Assuming a miracle happens overnight, and you feel (emotion from question before) towards cybersecurity - What would change?

    - What would have happened for you to now feel this emotion?

    - What emotion would you prefer not to feel towards cybersecurity?

    - What would have happened for you to now feel this emotion?

Consequences

- The following questions were asked:

    - Do these emotions have an impact on your behavior (Top emotions) towards cybersecurity? How?

    - How do your emotions towards cybersecurity influence your daily work/primary tasks?

Coping

- The following questions were asked:

    - Is there something that helps you deal with these emotions? What?

    - Is there something your company/employer can do to address these emotions? What?

Self-efficacy

- A scale was displayed in miro

    - How confident are you in your ability to engage with cybersecurity in general (e.g., learning cybersecurity content or implementing company guidelines)?

    - Why is that the case?

Interview Demographics

Participant	Age	Gender	Industry	Work experience (years)	Interview duration
P1	20 - 24	f	Research and education	1 - 5	0:45
P2	20 - 24	f	Research and education	1 - 5	0:40
P3	20 - 24	f	Marketing	1 - 5	0:43
P4	25 - 29	f	Finance	1 - 5	0:46
P5	20 - 24	m	Engineering	1 - 5	0:43
P6	50 - 54	m	Pharmaceuticals	21 - 25	0:42
P7	60 - 64	m	Engineering	36 - 40	1:07
P8	20 - 24	f	Research and education	1 - 5	0:24
P9	50 - 54	f	Healthcare	16 - 20	0:30
P10	20 - 24	m	Research and education	1 - 5	0:32
P11	20 - 24	f	Healthcare	1 - 5	0:30
P12	55 - 59	m	Information technology	31 - 35	0:35
P13	30 - 34	m	Consulting	11 - 15	0:40
P14	18 - 19	m	Healthcare	1 - 5	1:15
P15	25 - 29	m	Consulting	1 - 5	1:15
P16	35 - 39	m	Insurance	16 - 20	1:04
P17	45 - 49	m	Research and Education	16 - 20	1:19
P18	30 - 34	m	Public sector	6 - 10	1:12
P19	45 - 49	m	Information technology	21 - 25	0:48
P20	50 - 54	f	Administration	31 - 35	1:08
P21	25 - 29	f	Consulting	6 - 10	0:54
P22	55 -59	f	Research and education	21 - 25	1:27
P23	30 - 34	m	Administration	11 - 15	1:15
P24	30 - 34	m	Engineering	6 - 10	0:53
P25	35 - 39	m	Engineering	6 - 10	0:55
P26	25 - 29	f	Pet sector	1 - 5	0:53

Table 2: Participant demographics. For privacy, department and rank are omitted; industries, age and work experience categorized

Scale	Variable	M	SD	MIN	MAX	MEDIAN
SREIS	Perceiving Emotion	3.77	0.48	2.75	5.00	3.75
SREIS	Use of Emotion	3.10	0.75	1.00	4.33	3.00
SREIS	Understanding Emotion	3.23	0.72	2.00	5.00	3.25
SREIS	Managing Emotion (self)	3.46	0.71	2.00	4.75	3.50
SREIS	Social Management	3.68	0.59	2.50	4.75	3.75
SREIS	Emotional Intelligence Score	3.45	0.36	2.87	4.30	3.41
HAIS-Q	Knowledge_Password management	4.71	0.43	3.67	5.00	5.00
HAIS-Q	Knowledge_Email Use	4.26	0.62	2.67	5.00	4.33
HAIS-Q	Knowledge_Internet use	4.47	0.65	2.67	5.00	4.67
HAIS-Q	Attitude_Password management	4.71	0.40	3.33	5.00	4.83
HAIS-Q	Attitude_Email Use	4.56	0.43	3.67	5.00	4.67
HAIS-Q	Attitude_Internet use	4.63	0.43	3.67	5.00	4.67
HAIS-Q	Behavior_Password management	4.68	0.41	3.67	5.00	5.00
HAIS-Q	Behavior_Email Use	4.40	0.65	3.00	5.00	4.67
HAIS-Q	Behavior_Internet use	3.90	0.78	2.67	5.00	3.83
HAIS-Q	SUM_Password management	14.09	0.95	11.33	15.00	14.33
HAIS-Q	SUM_Email Use	13.22	1.45	9.67	15.00	13.33
HAIS-Q	SUM_Internet use	13.00	1.57	9.67	15.00	13.33
ISCI	ISCI_Practices	6.69	2.57	3.00	12.00	6.00
ISCI	ISCI_Importance	12.54	2.16	8.00	15.00	12.50
ISCI	ISCI_Laxness	5.04	1.97	3.00	9.00	4.50
ISCI	ISCI_Score	10.73	1.41	7.67	13.67	10.67

Table 3: Screening. Controls and variables to maximize variation. EI was measured to ensure emotions reflection skills. We retained all participants to preserve diversity and avoid bias, monitoring those with slightly noticeable scores without issues.

C   Appendix: Qualitative Survey

Qualitative Survey: Method

Welcome. Participants were provided information on the study's conditions, procedure, and purpose, including background details on participant rights and data processing, and granted consent upon agreement with the outlined conditions.

Emotional Cybersecurity Events, Emotions towards Cybersecurity and Consequences.

- When you think about cybersecurity at work, what emotions do you feel?

- Put yourself in these emotions. Why do you feel these emotions towards cybersecurity at the workplace? Are there specific events that led to these emotions?

- What was the result of these emotions? e.g. Do your feelings affect your security behavior or the way you approach your work? How does this affect your attitude toward work?

Thoughts on cybersecurity. Based on Renaud et al., participants were asked to describe their spontaneous thoughts about cybersecurity in open questions and to record what was unsaid [78].

- What are the first thoughts that come to mind when you hear the term of 'cybersecurity'?

- What have you always wanted to say about cybersecurity?

Cybersecurity definition and behavior. A brief definition of cybersecurity was introduced, and participants were asked to name behaviors they feel are necessary to protect cyberspace within organizations. Separately, participants were asked which measures they actually implement.

- What should you do to protect yourself against cyber attacks at the workplace?

- What measures do you actually take to protect yourself against cyber attacks at the work place?

Cybersecurity Incident Experience. Participants who could not name any experiences were allowed to skip the item.

- Have you ever been the victim of a cyber attack? Please describe your experience as detailed as possible. Place emphasis on your emotional journey throughout the experience.

Closing. Cybersecurity-specific, organization-specific and general demographic data was collected. To collect security-specific data, the Security behavior intentions scale (SeBIS; [36]) for the collection of behavioral intentions and the SA-6 for the collection of security attitudes [39]. In addition, information on gender, age, education, employment status , industry and company size were provided.

Qualitative Survey Demographics

Scale	Variable	M	SD
SeBis	Device Securement	4.39955357	0.66602819
SeBis	Password Generation	3.70758929	0.88065037
SeBis	Protective Awareness	3.9	0.87423436
SeBis	Updating	3.5922619	0.91689372
SA-6	Score	3.44494048	0.96192092
Age Group			
< 19		1	
20 -24		29	
25 - 29		22	
30 - 34		9	
35 - 39		11	
40 - 44		9	
45 - 49		5	
50 - 54		9	
55 - 59		14	
60 - 64		2	
Gender			
female		32	
male		78	
non-binary		2	
Company Size			
1-9		10	
10-49		18	
50-249		14	
250-1000		15	
>1000		54	
Industry			
Chemistry & Raw Materials		3	
Agriculture		1	
Construction		4	
Services & Crafts		3	
Energy & Environment		2	
Finance, Insurance & Real Estate		26	
Commerce		2	
Internet		4	
Consumption		1	
Media		4	
Metallurgy & Electronics		2	
Pharmaceuticals & Health		9	
Education		6	
Technology & Telecommunications		7	
Tourism & Hospitality		1	
Transportation & Logistics		2	
Economy & Politics		7	
Other		28	

Table 4: Participant demographics. Quantitative measurements were included to add further depth to the understanding of the sample and ensure a diverse representation across selected variables.