Exploring Authentication for Security-Sensitive Tasks on Smart Home Voice Assistants

Alexander Ponticello and Matthias Fassl, *CISPA Helmholtz Center for Information Security and Saarland University;* Katharina Krombholz, *CISPA Helmholtz Center for Information Security*

https://www.usenix.org/conference/soups2021/presentation/ponticello

This paper is included in the Proceedings of the Seventeenth Symposium on Usable Privacy and Security.

August 9–10, 2021

978-1-939133-25-0

Open access to the Proceedings of the Seventeenth Symposium on Usable Privacy and Security is sponsored by

Exploring Authentication for Security-Sensitive Tasks on Smart Home Voice Assistants

Alexander Ponticello
CISPA Helmholtz Center for Information Security and Saarland University

Matthias Fassl
CISPA Helmholtz Center for Information Security and Saarland University

Katharina Krombholz
CISPA Helmholtz Center for Information Security

Abstract

Smart home assistants such as Amazon Alexa and Google Home are primarily used for day-to-day tasks like checking the weather or controlling other IoT devices. Security-sensitive use cases such as online banking and voice-controlled door locks are already available and are expected to become more popular in the future.

However, the current state-of-the-art authentication for smart home assistants consists of users saying low-security PINs aloud, which does not meet the security requirements of security-sensitive tasks. Therefore, we explore the design space for future authentication mechanisms.

We conducted semi-structured interviews with $N = 16$ Alexa-users incorporating four high-risk scenarios. Using these scenarios, we explored perceived risks, mitigation strategies, and design-aspects to create secure experiences. Among other things, we found that participants are primarily concerned about eavesdropping bystanders, do not trust voice-based PINs, and would prefer trustworthy voice recognition. Our results also suggest that they have context-dependent (location and bystanders) requirements for smart home assistant authentication. Based on our findings, we construct design recommendations to inform the design of future authentication mechanisms.

1 Introduction

Voice-controlled smart home assistants find their way into more households every year. Gartner estimates that, by the year 2025, half of the knowledge workers will use voice assistants every day [10]. Currently, voice assistants offer entertainment (e.g., playing music, games), information gathering (e.g., weather, cooking recipes), and personal planning (e.g., calendar, task list). They are also a control hub for smart home IoT devices, such as smart light bulbs or heating. However, vendors already work towards new and more security-sensitive use cases for these assistants. Voice-based online shopping allows users to order goods without interrupting their current activity. Compatible locking systems permit users to open doors via voice commands [7]. Capital One, a technology-focused bank in the U.S., uses the Amazon Alexa platform to offer bank services such as retrieving account information, including their current balance, or paying credit card bills [12].

However, as Abdi et al. [2] found, security and privacy concerns hinder user adoption of these new use cases for voice assistants. Amazon Alexa, a widespread voice assistant that supports online shopping, currently only offers an optional voice code to authenticate users before their purchase. This simplistic authentication method is insufficient for more security- and privacy-critical tasks. Hence, voice assistants need more robust protection mechanisms. Our community already invested a significant effort in developing and improving authentication mechanisms for various tools and use cases [9, 14, 20]. However, designing authentication for voice assistants comes with unique challenges since they usually do not offer I/O methods beyond the voice channel. This limitation makes transferring existing authentication mechanisms to voice assistants difficult. Hence, we need device-appropriate authentication mechanisms for voice assistants. Developing these starts with finding all viable forms of authentication that users trust.

In this work, we explore the design space of authentication with voice assistants in a user-centered way. We conducted semi-structured interviews with $N = 16$ participants that included four scenarios. These scenarios depicted different situations in which the protagonists perform security-sensitive tasks with a voice assistant. We evaluated the transcribed in-

USENIX Symposium on Usable Privacy and Security (SOUPS) 2021.
August 8–10, 2021, Virtual Conference.

terviews using Thematic Analysis [11] to explore the design space. Our contribution includes findings on: (1) users' perception of threats, (2) users' mitigation strategies in security-sensitive circumstances, (3) users' expectations for authenticating with voice assistants, and (4) implications for the design of future authentication mechanisms. Our results show that users see bystanders in hearing range as a potential threat to their security and privacy. Their main mitigation responses focus on limiting their use of security-sensitive features. Hence, developing alternative user-trusted authentication mechanisms is crucial to facilitate adoption of security-sensitive use cases. The participants appreciated the low-effort interaction with voice assistants and expected similar from authentication. Voice-based biometric authentication fulfills that criterion and was frequently suggested for authentication. In social situations, participants reported discomfort with voice code authentication and privacy-sensitive tasks. Hence, an additional *discreet mode* for voice assistants potentially improves adoption rates. Participants described that their trust in security mechanisms builds with experience. Therefore, we suggest that voice assistants provide a *demonstration mode* for security- and privacy-related features.

2 Related Work

Our work builds on different areas of prior work: security and privacy of smart home environments as well as voice assistants, alternative authentication schemes for voice user interfaces, and users' risk perceptions and mitigation strategies.

2.1 Security and Privacy of Smart Homes

Zeng et al. [39] studied users' mental models of smart home systems and threats. They found incomplete mental models of how IoT devices, including voice assistants, interact with each other and with back-end cloud services. Building upon this work, Zeng and Roesner [40] explored users' security and privacy issues in a month-long in-home study. They identified several open challenges, most importantly incorporating voice assistants into access control systems so that they can become effective control hubs for smart homes. In this context, they highlight the importance of sophisticated voice-based authentication, which motivates our study.

Yao et al. [37] conducted a co-design study with users and non-users of smart home technologies to investigate their privacy concerns and needs. They identified key design factors for smart home privacy controls, including authentication for multiple users and access control.

Zimmermann et al. [43] studied potential users' mental models of smart homes. Their participants had sparse mental models of smart home systems, and almost all of them were concerned about their personal data's security.

Yao et al. [38] used three scenarios to study bystanders' privacy perceptions in smart homes, i.e., people living in or visiting smart homes where they are not primary users. Bystanders were concerned about the video and audio data collection and demanded privacy controls tailored to them specifically. Furthermore, the authors highlight how the users' role in smart homes, e.g., system owners and bystanders, can strain their relationship.

2.2 Security and Privacy of Voice Assistants

Huang et al. [18] examined privacy perceptions and coping strategies of users sharing voice assistants. They found that users with limited mental models did not understand how the system shares their data with other users. In contrast, participants with more advanced mental models were concerned about the immature technology, e.g., voice recognition to distinguish users. The authors highlight the need for more sophisticated authentication mechanisms to tackle these issues.

Lau et al. [21] conducted a diary study and interviews to shed light on privacy perceptions and privacy-seeking behaviors around voice assistants. They found that voice assistant users did not entirely understand privacy risks and frequently traded privacy for increased convenience. Non-users were concerned about privacy and security, partially because of their limited trust in voice assistants' manufacturers.

Chalhoub and Flechais [13] report similar findings from their qualitative study exploring the effect of user experience (UX) factors on voice assistant users' security and privacy. They found that common security and privacy features, such as muting, were not user-friendly. As a response, users disabled features or disconnected their devices.

Zhang et al. [41] describe the *Dolphin Attack*, a novel technique that utilizes ultrasonic audio signals to inject commands into smart home voice assistants. These commands are inaudible to humans and exploit the non-linearity property of current microphones, which is why they treat high-frequency sounds similar to genuine human speech. The authors bypassed the biometric voice authentication of up-to-date smart home voice assistants by combining this attack with users' resampled legitimate audio snippets. Roy et al. [29] build upon this work, extending the attack range from 1.5m to 7.6m using an array of speakers. Sugawara et al. [34] developed an attack called *LightCommands*. By exploiting a vulnerability in micro-electro-mechanical systems (MEMS) microphones, this technique allows an attacker to inject commands via a potent light source. Since these commands are transmitted by light, attackers can inject them from afar while victims cannot hear them. The authors report a successful command injection over a distance of 75m with a laser beam aimed at a Google Home device behind a glass window. Lei et al. [22] make use of channel state information in Wi-Fi networks to detect human presence in a room. Assuming that most attacks happen during users' absence, VUI systems only ac-

cept commands if someone is present at that time. Related work identified a gap between users' expectations of potential threats and technically feasible attacks. Using our study, we also want to increase our knowledge about this gap and laying the groundwork to reduce it in the future.

2.3 Authentication for Voice User Interfaces (VUIs)

Feng et al. [14] designed a wearable-based authentication scheme for VUI. Their system verifies VUI commands by independently recording voice commands from skin vibrations. Hence, this method provides continuous authentication. They tested different designs for the wearable, such as earbuds, necklaces, and glasses.

Blue et al. [9] proposed a similar scheme using a second microphone-equipped device, e.g., a smartphone. By measuring the direction of arrival of each voice command, their system can detect whether the speaker is closer to the VUI or the second microphone. Assuming that users carry their smartphone on them during VUI interaction, the system only deems nearby commands authentic.

Kwak et al. [20] employed machine learning to differentiate genuine user commands from malicious input.

Zhang et al. [42] developed a system for speaker liveness detection. By extracting features in the Doppler shifts, they can distinguish audio generated by an artificial speaker from a human voice. Their system enhances voice authentication by protecting from common threats, e.g., *Replay Attacks*.

The presented authentication systems build upon assumptions about how users interact with VUIs and how they perceive the system. In this work, we use qualitative methods to explore the underlying design space, thereby laying the groundwork for future authentication systems taking users' security and privacy needs into account.

2.4 Users' Risk Perceptions and Mitigation Strategies

Several other works used methodological approaches similar to this paper's to investigate users' risk perceptions and mitigation strategies outside of the smart home context. Many of their findings are observable across various systems and technologies. Hence, they potentially apply to voice assistants as well.

Harbach et al. [16] studied Internet users' risk awareness. The results indicate that most of the 210 participants were aware of general risks. The authors state that users are aware of seven risks on average, which significantly vary across persons, populations, and the interaction's context. They highlight that existing security measures often focus on technical risks of which users are less aware. Since users have a limited compliance budget, the authors argue that they might not adopt measures that do not directly address their perceived relevant risks. Furthermore, the authors propose improving risk communication and education to support the users' risk perception.

Ruoti et al. [30] conducted interviews with middle-aged suburban parents about their online security posture. They found that users weigh the trade-offs between gained security and necessary effort when choosing security mechanisms. Due to participants' perception that complete security is unobtainable, they less frequently adopt cost-intensive coping strategies. They identified a four-step process users pass through where they first learn about a new security threat (e.g., by news reports), evaluate their personal risk (ignoring threats perceived as unlikely), estimate the damage in terms of the effort they have to invest after a breach, and selecting an appropriate coping strategy after weighing the costs and benefits.

Stobert and Biddle [33] studied coping strategies that users apply when managing passwords. They found that some of the most prominent mitigation strategies, e.g., writing down passwords or reusing them, seem to disregard popular security advice. The authors argue that this behavior is not caused by users' insufficient risk perceptions but rather that users make rational choices based on their personal resources.

3 Methodology

We chose semi-structured interviews building on previous works [8, 39] to answer the following research questions:

RQ1 Which attackers and threats are users concerned about when performing high-risk tasks via voice-controlled assistants in a smart home environment?

RQ2 Which potential mitigation strategies do users apply to protect themselves?

RQ3 Which properties does an authentication system for voice assistants need such that users perceive it as secure?

3.1 Procedure

We briefed participants on the topic and purpose of the study and how data is processed and handled. All participants signed consent forms that permitted audio recordings. Our interview guideline (presented in Section A) consists of three parts. First, we asked a series of warm-up questions regarding our participants' general Alexa usage and experiences with the online shopping feature.

In the second part of our interview guideline, we presented participants with four scenarios on *vignettes*. These included pictures and short textual descriptions of an interaction between a user and their smart home assistant. Prior work showed that scenarios are a useful tool for examining users'

perceptions and mental models of a system [2, 4, 19, 36]. Vignettes allowed participants to immerse themselves into situations, which would have been more difficult using interview questions alone. Vignettes are closer to reality than abstract questions and might reduce social desirability bias by allowing interviewers to ask questions less directly [26].

Scenarios The scenarios combined four security-sensitive tasks with different situations. These situations vary in two aspects: the number of bystanders and the location, namely inside or outside the house. Most of the presented functions are currently not available in central Europe. The scenarios are as follows:

- **Dinner.** This scenario combined the task of transferring a small amount of money during a dinner party with friends. The use of Alexa can be convenient since the user is sitting at a table. Several bystanders might eavesdrop on the interaction. However, these people are, to a certain degree, trustworthy as they are close acquaintances. The corresponding image shows a laid table with several people around, chatting in a light atmosphere.
- **TV.** This scenario involves users and their partners. We selected the activity of paying a reoccurring bill since this is a typical task concerning both partners while also being less casual and less frequent. The picture associated with this scenario depicts two people sitting in a living room on a couch in front of a running TV.
- **Door.** We combined the task of unlocking the front door with the scenario of coming back from grocery shopping. A typical task that users perform while outside the house. The situation includes the user carrying several bags, making unlocking doors more difficult. This setting justifies the use of a voice-controlled smart home assistant. No other people are immediately present in the scene. The picture shows a person carrying bags of groceries next to a car, a blue sky in the background indicates that the scene takes place outside.
- **Hands.** We coupled the task of checking a transaction history with gardening work, making the protagonist's hands dirty. We included children as potential bystanders. We described them as running around and screaming, meaning they do not pay immediate attention to the user while still being present. Also, this scenario does not feature a dialog with Amazon Alexa. In this description, we do not refer to Alexa nor include a device in the image to leave room for the interviewee to imagine how an interaction could play out. At the same time, this allows us to explore alternative interaction mechanisms, potentially not involving Alexa.

We presented the scenarios in random order. During on-site interviews, we presented the vignettes on printed and laminated cards face-down to participants. For remote interviews, we showed participants a website that displayed four face-down cards. In both cases, we flipped and discussed the cards in the participants' chosen order. Section C presents the full vignettes that we used for the interviews. We provided pen and paper for note-taking to participants or asked them to send us their drawings by email during remote interviews respectively.

After letting them read the description text and look at the image, we asked participants which problems they think could arise in such a situation. We did not ask about security-related problems to avoid priming participants in a specific direction. If participants mentioned no security-related problems, we followed up with respective questions, e.g., whether they thought the voice code included in the scenario was useful or not. Then, we explored threats that participants identified and asked them to think of any other actors posing threats and their potential mitigation strategies. We investigated what interviewees thought might be useful to them and which mitigation strategies they would apply in the given scenario, with the threats described above in mind. We repeated this process for all four scenarios. Afterward, we asked participants to summarize all four situations and to think about possible similarities and differences between the scenarios, possibly applying the insights they gained in a later scenario to an earlier one. This recapitulation also helps to focus participants on details they might not have noticed before (e.g., the number of people present in the scenario) and think about the consequences introduced by said factors.

In the third and final part of our interview guideline, we included some demographic questions, mostly used to describe our sample. We included two standardized scales in this section, namely the ATI scale [15] and the CFIP scale [32].

After each interview, we asked the participant about any remaining questions, reiterated the study's purpose, and explained why we designed the interview guideline and the vignettes as presented. We also explained the current situation regarding security, and most of all, authentication, on Amazon Alexa and comparable VUIs.

Accessibility One blind Alexa user participated in our study. We adapted our study material to ensure accessibility and exchanged the printed vignette cards with two separate audio recordings: To have a clear distinction between vignette descriptions and interview questions, one author, who was not the interviewer, narrated the picture displayed on top of the card. In a separate audio file, the narrator read the corresponding text aloud. We used a computer-generated voice similar to the one from Alexa to illustrate interactions with the voice assistant. Providing two recordings allowed the participant to replay each part separately.

Pilot Interviews We pilot-tested our interview guideline with two on-site and two remote interviews. Based on the

findings, we decided how much time to allocate as well as financial compensation. We dropped one interview question as it was too ambiguous; we also modified the presentation of the *door* vignette to clarify the scenario. We excluded the pilot interviews from the final dataset.

3.2 Thematic Analysis

We transcribed the data at an orthographic level, including non-verbal utterances only when we deemed them essential for the semantic of a phrase (e.g., a participant laughing while saying something, indicating it was a joke). Afterward, we read and re-read the data to get an even better understanding, taking notes of interesting details and basic patterns.

We chose to analyze our data using a thematic analysis approach, as described by Braun and Clarke [11]. We conducted the interviews in English and German, coded the resulting data in German, and later translated the codebook to English.

To construct a codebook, two researchers performed open and axial coding on a subset of four interviews. First, we performed open coding, then met to resolve disagreements and re-coded the data. Krippendorff's alpha was 0.50 before and 0.94 after the discussion and re-coding step, indicating a high agreement. Then, we performed axial coding (on the same subset of interviews) to identify higher-level themes. Then, another subset of four interviews was coded with a Krippendorff's alpha of 0.83, indicating a strong agreement between the two coders. At this stage, the existing codebook covered most of the data's aspects, so we only sparingly introduced new codes. Finally, one researcher coded the remaining interviews using the codebook agreed upon in the previous discussion.

3.3 Recruitment and Participants

We mainly recruited Alexa users because Alexa has the largest share of the smart speaker market, and its (security-sensitive) shopping feature is well-developed and widespread. However, we also welcomed participants who had experience with other types of voice assistants.

In total, we recruited 16 participants in Germany (9), Austria (6), and Italy (1); five of them via flyers around our institution's campus, three participants over mailing lists; six via convenience sampling; and two via snowball sampling. We stopped recruiting new participants after we reached saturation for our target population, i.e., Amazon Alexa users from Central Europe without computer science background. Due to the COVID-19 pandemic, we conducted ten interviews online and six in person at our department. We compensated all participants with a 15 Euro Amazon voucher, which is in line with similar studies [19, 39].

In total, we recruited seven women and nine men. Their average age was 29.31 ($\sigma = 10.69$, median $= 26.5$). Fourteen participants had at least completed high school, with seven holding a bachelor's degree (or equivalent), and two holding a master's degree (or equivalent). We also measured the participants' affinity for technology interaction using the seven-point ATI scale [1]. The average ATI score was 4.1 ($\sigma = 0.76$, median $= 3.83$), which is above the population-wide average of 3.5. To assess people's privacy concerns we used the seven-point CFIP scale [17]. The average CFIP score was 5.76 ($\sigma = 0.71$, median $= 5.93$).

3.4 Ethical Considerations

Our institution's ethical review board (ERB) reviewed and approved our study. We followed our principle of minimizing the collection of personally identifiable information (PII) as far as possible. We stored and processed data in line with the GDPR and our institution's ethical regulations. We collected informed consent from all participants and informed them how we would process their data. If participants had further questions or wished to withdraw their consent afterward, they could use the provided contact information.

3.5 Positionality Statement and Expectations

In the spirit of constructivism, we assume that our personal views as researchers shape every part of a study, from study design to data analysis to reporting. Here, we want to make our a priori expectations (similar to Krombholz et al. [19] and Braun and Clarke [11]) transparent. We focus on the expectations that influenced the design of the four scenarios.

We expect that the presence of bystanders (esp. considering the familiarity between the user and the bystander), the location in which users perform a task, and the task's perceived security-sensitivity (esp. considering financial risks or potential risk of a property's physical security) are most likely to influence the participants' responses.

E.g., we expect users to neglect the threat of other IoT devices listening in on their actions but hypothesize that they perceive bystanders as potential risks. Furthermore, we expect that users have incorrect assumptions about the security of authentication methods and the kind of threats they mitigate. Regarding mitigation strategies that users employ, we expect to find that people refrain from using the system entirely or only use security-critical features when bystanders are not present. Some users might use Alexa's whisper mode to prevent other people from overhearing a sensitive conversation.

4 Results

We now present the exploratory findings of the design space for smart home assistant authentication. When analyzing our data, we focused on answering our research questions stated in Section 3. This chapter is structured according to the categories we developed during the axial coding step. First, we cover the perceptions of threats. Users were concerned about

different attackers that could affect them in the presented scenarios. They also reflected on trust in certain groups of people or entities. Next, we report mitigation strategies that participants considered to protect themselves. These mitigation strategies improve our understanding of how participants use these systems and which practices they adopt to mitigate threats. Finally, we present essential properties for secure and usable smart home assistant authentication we discovered during our data analysis.

For easier readability, we refer to individual participants with labels P1-16 throughout this section.

4.1 Concerns about Attackers and Threats

We answer RQ1 by reporting perceptions of threats and attackers that users were concerned about when performing security-sensitive tasks on a voice assistant. We found that most users perceived bystanders as potential threats. Both familiar (e.g., family, friends) and less familiar (e.g., neighbors, casual visitors) bystanders could be present during an interaction with Alexa, meaning that the voice code used for authentication could be eavesdropped on by an intentional attacker or an accidental listener.

Insiders We discovered several conflicting perceptions about insiders as a threat. Similar to previous work [18, 24], almost all participants agreed that they trust their friends in general, however, we found that this does not always extend to security- and privacy-related affairs. P7 states: *"I trust my friends, but not with my money."* Correspondingly, most interviewees showed a more extensive amount of trust towards a partner, some of them even willingly sharing their authentication code. Others, however, expressed concerns that a partner might become a threat if the relationship were to end on bad terms. Previous work by Levy et al. [23] and Marques et al. [25] suggests widespread adversarial behavior between family members. Lastly, we found the perception that children are a potential threat, depending on their age. P4 explains that: *"Children are usually quite bright and soak everything up like a sponge, and I think they could use that somehow, the voice code, to make transfers or top up their phone."*

When we asked participants about the possible motivation of insider threat actors, they suspected that friends and children would prank them. While such pranks usually do not cause much harm, they present an inconvenience that most participants prefer to avoid.

Criminals Participants also considered more serious attackers such as criminals, both on- and offline, which is in line with results of previous work [2, 16, 18, 30, 43]. In the presented scenarios, criminals could be especially motivated by the potential high financial gains. As also reported by other researchers [39, 43], physical access to their home proved to be a primary and widespread protection measure within our sample. In our specific scenarios, interviewees showed awareness of several attack vectors, namely eavesdropping, *Replay Attacks*, and brute-force attacks on voice codes. Participants expected attackers to employ readily available devices such as microphones to capture a user's interaction with a voice assistant. While most thought such an attack would need to happen in situ, a few interviewees were aware of voice sampling techniques using arbitrary audio of a user to produce adversarial samples.

In the context of personal finance scenarios, participants were concerned about remote attackers interfering with their devices over the Internet. These attackers could exploit Alexa's vulnerabilities to eavesdrop on a user's voice code or inject malicious commands directly, in both cases bypassing authentication. Some interviewees also suspected that attackers use other IoT devices to monitor users and interfere with voice assistants. Similarly, some were aware of malicious skills as potential attack vector.

Untrustworthy or faulty infrastructure Due to past experiences, interviewees expressed concerns about technical issues impeding a secure interaction with Alexa. They highlight that failures of the speech-to-text system might lead to wrong or unauthorized commands getting executed, marking a security breach. These findings add to results by related work, demonstrating how unintended or miss-interpreted voice commands can lead to violations of users' privacy [24] and frustration during authentication [35]. Schönherr et al. [31] found over 1000 triggers for Amazon Alexa, Google Assistant, and other smart home assistants in TV-shows, news, or audiobooks.

Finally, almost all users expressed privacy concerns when it comes to sharing data with Amazon. High-risk tasks such as money transfers can involve sensitive data that participants were uncomfortable sharing with a company they suspected of employing targeted advertisement or selling data to third parties. Storing user data renders data leaks on the backend of the system possible, potentially due to cyberattacks. Finally, some users also explained that Amazon or its employees might eavesdrop on a user's voice code and use it against their will.

4.2 Mitigation Strategies

We address RQ2 by reporting the participants' mitigation strategies. Users largely agreed they would refrain from using an authentication system they perceive as insecure, especially if they consider the use case non-essential. This matches users' coping strategies in other contexts, e.g., online shopping or setting up smart home systems [2, 18, 24, 39]. Our findings suggest that users generally perceive Alexa as a luxury item that facilitates tasks but does not enable previously unavailable features. Hence, using Alexa for security-sensitive tasks is just an additional attack surface for the participants. As P4

phrases it: *"I wouldn't use any of the skills described here because the effort- or the comfort-to-risk ratio is not profitable for me."* We found, similiar to Abdi et al. [2], that users preferred employing personal computers or smartphones as fall-back authentication method. Mainly because users have pre-established trust with these devices.

We found that users employ a trial-and-error strategy to build up trust and improve their understanding of the protection provided by authentication systems. By trying out the system under typical attacking conditions, users could gain trust in a novel mechanism. We found a go-to attack for this technique is mimicking a legitimate user's voice. Users would test voice biometric authentication by *"sit[ting] in front of it quite often and try[ing] it out while disguising my voice, to see whether Alexa still recognizes me or not."* (P10). Most interviewees had not used voice-based authentication before and did not trust a system without hands-on experience. We found that both positive past experiences and a lack of negative ones can give users a sense of security. P1 explains this as follows: *"there may be some [security] issues with the payment method I'm currently using, but I've done it so often and I'm so familiar with it that I feel safer because of that."*

We found that eavesdropping was the users' prime concern. Hence, interviewees presented various mitigation strategies for this threat. The most prominent one was moving to another room if several bystanders were present, e.g., in the scenario *"Dinner"*. Huang et al. [18] reported similar user concerns and coping mechanisms in a less security-sensitive task: making phone calls via voice assistants. We found that there exist specific situations in which users do not desire voice interaction. Some participants stated that this was due to an awkward feeling when talking to a computer, which can be perceived as *"admitting to being lazy"* (P10) because a user does not carry out tasks themselves, delegating them to a computer instead. Furthermore, interaction over voice can draw unwanted attention to the user. Participants expressed a desire for discreet interaction options, especially for money-related tasks; *"money is always a delicate topic and you don't want to address that in front of everyone"* (P4). Using the whisper mode of Alexa can be a less obtrusive operation mode. Participants also stated that this mode potentially mitigates eavesdropping. However, this input feature was perceived as less elegant and, consequently, not fitting into *"the Alexa lifestyle"* (P6).

Another mitigation strategy for eavesdropping was changing the code regularly. By doing so, participants expected that a leaked code would no longer be valid during an attack. Similarly, interviewees described more complex codes as hard to remember and, therefore, also difficult for an eavesdropper to pick-up. We found that users believed they could recognize on-going attacks against their devices while present. P7 notes: *"Inside the house, no real sound can get through. If someone stands in your garden and yells: ALEXA! [. . .] then you probably hear it too."* Therefore, attacks would mainly occur while they were away from home. In this case, participants desired stronger than usual security measures. Our findings suggest that users are not aware of attacks injecting inaudible voice commands, possibly from outside the house [34, 41].

While participants did not perceive voice codes as an adequate authentication mechanism for general use cases, some interviewees talked about its positive effects. A voice code can be an effective mitigation strategy against accidentally executed commands since, unlike regular voice commands, participants did not imagine saying their code in a casual conversation. Some interviewees perceived the code as a minimum security mechanism protecting them from their friends' or children's pranks. They preferred using a code over having no security measures. P9 explains that it *"just gives another layer of security, so my friend couldn't just come into my house and be like: Alexa, pay the utility bill!"* Several participants mentioned remote attackers as a concern, though none could think about mitigation strategies against this threat. Participants did not talk about preventive measures such as keeping systems up to date during the interviews. This observation is in line with Anell et al.'s findings [6].

4.3 Important Properties of Authentication Systems

We present important aspects of authentication systems for voice assistants we identified to address RQ3. These properties showed to be crucial for users' perception of security when performing security-sensitive tasks.

Building Trust

Our participants' perception of security in the context of sensitive tasks on voice assistants was tightly couple with trust in the system. This matches findings about privacy perceptions in shared-user settings by Huang et al. [18] those of bystanders in smart homes by Yao et al. [38]. Participants did not trust a new system out-of-the-box. However, they described several ways to establish trust, especially towards an authentication system. One reoccurring theme was that users transferred trust from a trusted entity to a new system it supports. Interviewees named mostly banks as an example of such an entity, but also *PayPal* and energy providers. Participants stated they would trust a system more if a trusted third party provided it directly. In the words of P8: *"So if it really came from the bank, I'd trust the whole thing more, then I'd be more inclined to use it."* Users apply past experiences to root their trust in entities and are convinced of these entities' interest in keeping their systems secure.

We furthermore found that participants who describe themselves as *"old school"* (P4) were skeptical of novel systems and perceived themselves as less likely to adopt them. Interviewees expected younger users to have an easier time adjusting to a new system. As P10 states: *"It is not normal*

for my mom to do banking on her phone. [. . .] It will perhaps be normal for the next generation to tell Alexa such things."

Positive experiences with a system in the past led to a higher trust in its security. Similarly, users could lose trust by witnessing security incidents. Applying a trial-and-error strategy to authentication can facilitate experiencing a system in a shorter period. Similarly, users could establish trust by checking other users' reviews and ratings. Reading about other people's experiences can have a similar effect on users' trust as experiencing something first-hand. P1 notes: *"If you read that everything works, you have many people who rated this if the reviews are consistently positive, that would certainly build up trust."*

Transparency and Agency

Almost all participants stated that transparency is essential when it comes to the perception of security. A transparent system can enable users to make informed decisions when interacting with such devices. Several participants noted that this property did not transfer well from computers or smartphones to smart home assistants. This attitude was partially due to the fact that voice rendering is difficult to understand for users as the underlying technical fundamentals and information-sharing models are complex.Visual interaction enables users to grasp information much quicker, as stated by P7: *"When I order on the PC, I have several options that I can grasp directly and it is simply easier for me to take in with my eyes than to listen with concentration."* This confirms Abdi et al. [2] who found that visual interaction enables users to absorb information more easily when shopping online.

Using a computer also conveyed a feeling of being in control, which we found is an important characteristic when it comes to security-sensitive tasks. Our findings suggest that using Alexa, in contrast, is perceived as surrendering agency over to another party. Users no longer perceived themselves as the active part and could only hope for the successful execution of the process. They attributed this feeling to an intransparent control flow. P10 states: *"It's weird if I don't see when something happens. Because I say something and it happens. And then I just can't understand whether it was done correctly."*

Our results suggest that the personification of Alexa is a potential factor for this perceived loss of agency. Interviewees compared Alexa to a human operator and expressed that voice commands felt like giving orders to an employee. This perception entailed that Alexa could be affected by human error. P7 explains: *"Suppose I had a butler and I always had to tell the butler: open the front door. I can't trust that* 100% *either. Clearly, somehow, there is a large basic trust. But even then it's kind of uncomfortable when you have in the back of your mind: what if he didn't do it, what if he forgot about it?"* Participants wished for a more transparent control flow, which could lead to an improved understanding of involved entities and task distribution within a system. Voice assistants could accommodate for this by explicitly stating control switches to the back end or third-party services.

Risk Assessment of Authentication

Users' assessment of risks proved to be an important factor in various contexts [18, 21, 30, 39]. Based on their personal assessment, our participants derived variable requirements for an authentication system. We identified an interaction's location as a major factor. Similar to Yao et al. [37], our findings suggest that some locations call for stronger security measures. Most participants agreed that the most distinctive difference was between interactions occurring in a public space (e.g., in front of the door) and those taking place in a private space (e.g., the user's home). Interactions in locations perceived as secure could use weaker authentication mechanisms. P9 states: *"If you're inside the house [. . .] I believe the voice recognition and the code would suffice plenty."*

Some interviewees also distinguished between different zones inside a home. Security-sensitive functionality could be limited to more private areas such as an office or a bedroom. P3 notes: *"Transactions are only allowed from the study, while for the device hanging in the children's room, or in the hallway area where everyone has access, only certain things work there."* Another factor of the risk assessment is being at home vs being away. In accordance with previous work [37], our participants perceived the threat of security breaches to be more prominent while they were away from home. Authentication systems could follow this assessment and apply stronger methods during the vacancy period.

Ruoti et al. [30] highlight how users weigh the perceived risk against the effort needed to protect their privacy. Similarly, we found that participants were comfortable with using weaker authentication mechanisms, if they considered an interaction to be low-risk. Several participants stated that they would prefer having no voice code when checking transactions. In contrast, most participants agreed that an authentication step should be in place to execute transactions. P14 explains: *"I would be fine with using a voice code to see my transaction history, even my account balance, [. . .] but to make a transaction, I don't think Alexa should be allowed to do that."* Some participants also expressed having different requirements of protection depending on the amount of money transferred. Low amounts could be sent without strong authentication. Finally, a few participants explained that, since absolute security did not exist, there has to be a trade-off. *"It just always depends on how much effort I want to put into it, there will be no absolute privacy with such a system."* (P16)

Perception of Authentication Methods

We structured our participants' insights on VUI authentication according to the following four authentication paradigms.

Knowledge-Based Authentication Participants thought of the voice code as a low-level barrier that could primarily mitigate casual attacks and pranks by familiar people. Similar to classic knowledge-based authentication, interviewees were concerned about confusing or forgetting the voice codes for different skills. We found that users could, therefore, revert to code reuse, also across platforms. P9 notes: *"I guarantee you if you have the voice code to open your front door that's gonna be your four-digit PIN for your debit card, it could be for plenty of things in your life."* Similarly, some participants described they would apply coping mechanisms transferred from passwords, e.g., modifying only the last digit between codes. This behavior entails potentially drastic consequences for voice code leaks as they could compromise the security of other systems as well. One participant stated that, while the voice code was not an acceptable authentication method for high-risk tasks, it could serve as duress mitigation. By setting up a code for threatening situations, a user could say that code instead of their usual authentication code, upon which the system would initiate an emergency routine.

Possession-Based Authentication Several participants stated that they would favor token-based authentication with Alexa. Tokens would not be susceptible to the openness of the voice input channel. Hardware tokens could detect the users' physical presence, which should match the Alexa device's location. A close-by token would then lead to the assumption that a legitimate user issued the voice command. P2 gives an example of using a smartphone as a token: *"Alexa can connect to my mobile phone, it's in the same location as I am communicating, then I guess it's fine."* P9 suggested that a microphone-equipped hardware token, such as a *"Fitbit"*, could be used as an authentication token. Devices carried by the user could be marked as trusted, which allows for weaker authentication mechanisms.

Another way how authentication with Alexa could facilitate smartphones would be push-notifications. Some participants expressed that getting a notification requiring confirmation whenever a security-sensitive voice command was executed could be a secure authentication mechanism. Similarly, OTP devices could replace a static voice code. In contrast, some participants stated that using an additional device for authentication would be *"defeating the point of the Alexa, being able to talk to a virtual assistant, now that you have to involve physical things to actually pay, so at that point, you just log into your phone and do it."* (P9)

Biometric Authentication We found that most users preferred biometric authentication due to the natural and effortless interaction with them. However, interviewees expressed concerns regarding the current state of voice recognition on Alexa. P16 states: *"It recognizes you by your voice, but this recognition sometimes doesn't work, and I think that's very rudimentary."* As some participants were aware of possible *Replay Attacks*, they expected future voice biometrics to distinguish live human speech from machine emitted sounds. Interviewees highlighted annoyance caused by false negatives as another drawback of voice recognition. Most participants reported past experiences where voice recognition did not function as expected, possibly due to natural variances in a user's voice. P12 explains: *The voice is often different, let's say when you have a cold, for example. Voice sounds different in the morning than in the evening.* In the context of the scenario *"Door"*, some users also brought up face recognition as a potential authentication mechanism used in combination with a smart home assistant.

Multi-Factor Authentication Some participants proposed combining some of the above-described methods to form stronger multi-factor authentication. Participants perceived that there is a direct relationship between more authentication factors and better security. We found that the preferred combination of authentication methods amongst participants is knowledge-based passcodes with voice biometrics. Other well-known high-risk systems that employ multi-factor authentication, such as bank accounts, probably influenced users' perception.

5 Discussion and Implications for Design

We discuss our main findings (i.e., the themes we identified during the analysis) along with our recommendations for design. We focus on aspects that were perceived as crucial for participants to feel protected during security-sensitive tasks.

Voice Recognition as an Intuitive and Trustworthy Authentication Method

In accordance with previous work [2], our participants found that voice recognition was the most convenient authentication mechanism for voice assistants. It was perceived as a natural way of authentication, as it resembles the human approach to identifying a familiar person, for instance, when talking on the phone. Complementary to known results, we found that this also holds when users perform high-sensitive tasks such as online-banking. Some smart home assistants currently employ a form of voice recognition to distinguish users. However, manufacturers, such as Amazon, do not yet recommend it as an authentication mechanism [5]. Participants were aware of potential shortcomings of voice recognition that researchers and developers need to address before users trust such a system. The most prominently expected feature was liveness detection which distinguishes human voices from speaker playback.

Users want to Test and Experience the Effectiveness of the Authentication Method

We observed that users initially mistrust new authentication mechanisms they had not used before. Some users tried to mimic other users' voice to *test* voice-based authentication. For novel biometric authentication schemes, we recommend including a *demonstration mode* which participants can use to try out the authentication process. Most state-of-the-art systems will block access once a user reaches a threshold of unsuccessful authentication attempts. Such systems are, therefore, not suitable for users to test different adversarial techniques. By including a separate sand-boxed mode that allows unlimited authentication attempts, users might build trust faster and understand novel interaction mechanisms better. Any such demonstration mode must have the same look-and-feel as the standard authentication process, the only difference being that upon successfully authenticating, no real user data is accessible. In this mode, the system should still inform users whether their authentication attempt was successful or not. Reynolds et al. [27] suggested a similar demonstration mode allowing users to verify the functionality of 2FA-tokens immediately after setup.

Users Want Unobtrusive Authentication for Social Situations

We found that participants felt uncomfortable using conspicuous authentication mechanisms in certain social situations. Hence, designs of authentication mechanisms for tasks in social settings need a *discreet mode*. This mode would replace the regular authentication mechanism with an unobtrusive alternative, allowing users to perform security-sensitive tasks without drawing attention to them. While conventional voice recognition has shown to be a desirable option, it might not work for settings that include several bystanders. Situations with considerable background noise make voice recognition inconspicuous, which, however, impedes the correct functioning of the smart home assistant's speech-to-text system, leading to failed authentication attempts. Participants reported having experienced such erroneous behavior before. An implementation of a new system could also automatically identify the current social situation a user is a part of during authentication by, e.g., detecting other persons nearby or measuring the level of background noise. The system could then dynamically adapt the authentication process according to predefined rules for different situations.

Low-Effort Interactions

We identified that effortless and straightforward user interaction are crucial adoption factors. Users reported that their main reason for using a smart home assistant was the low effort interaction with these devices, compared to computers or smartphones. If novel authentication mechanisms diminish the benefit of voice interaction by requiring interaction with other devices, users were no longer willing to use them since the perceived additional risk outweighed the benefits. Also, participants felt that if authentication with a smart home assistant required interaction with a smartphone, they could use the smartphone to perform the task instead. Therefore, the design of new authentication systems for use cases that are already possible with conventional platforms has to consider this risk-benefit analysis made by the users and reduce the effort needed to authenticate to an adequate amount. Such low-effort interaction could be provided by continuous authentication mechanisms, as described, e.g., by Feng et al. [14].

Transparent Authentication Processes

We found that participants were unsure about the information flow of an authentication process. Previous work [2, 38] suggests this is also the case for the general flow of privacy-related data in voice assistant ecosystems. In particular, which party performed the verification of the presented authentication information in scenarios involving third parties (e.g., banks) was not clear to all users. While some believed Amazon would authenticate the user and then get permission to access their account, others perceived Alexa as a literal assistant that takes a user's credentials and uses them to log into an application on the user's behalf. Two factors reinforce the users' perception that Alexa uses third-party systems in the same way a human user would: Alexa's output does not explain whether it came from Amazon or a third party, and users attribute human characteristics to conversational agents. To enhance transparency and make control flow transfers from the Alexa back-end to third-party skills easier to detect, we propose using different voices for each subsystem. This way, a user could instantly notice once the third-party takes over, resolving the aforementioned uncertainty. A similar mechanism to provide transparency could be having Alexa announce handing over control to a skill and reporting back once a request has gone through. This practice could improve users' understanding of the data flow and, consequently, result in more informed security decisions.

Account for Varying Requirements

In line with previous work [21, 38], we found that users have varying security and privacy requirements. In the authentication settings we studied, the two main factors were location and bystanders. In contrast to interactions inside the home, users were concerned about more threats for outside scenarios. In general, users were confident that they could detect malicious behavior from nearby bystanders. Therefore, fewer threats were relevant for such circumstances. Also, the security-sensitivity of the performed task affected the users' security requirements. Most participants agreed that information requests were less security-sensitive compared to tasks

involving money or physical access.

If the principal user was away from home, the smart home assistant should still be accessible or remain turned on. However, security mechanisms should become more restrictive, especially when it comes to authentication. A possible feature accounting for these varying requirements could be a guard mode that, if turned on, requires stronger authentication to turn back off. A real-life example would be an alarm system that only the correct code can disarm. A user could turn on the guard mode if they leave the house or go to bed at night. Upon their return, they authenticate once using a strong and perhaps a multi-factor authentication mechanism to turn guard mode off and switch back to the default authentication method, which could be weaker and less intrusive.

5.1 Limitations

Our sample included almost equal numbers of men and women. We also managed to recruit participants with a variety of educational backgrounds. However, the age distribution of participants skewed towards younger participants. I.e., our study underrepresents older users of smart home assistants. Our sample participants score slightly above average [15] when it comes to the affinity for technology interaction (ATI). CFIP scores indicate that our participants were highly concerned about their information privacy, indicating a further potential under-representation [28]. As this study is exploratory, we targeted users who already have experience with using Alexa. We recruited participants from Central Europe, in part via convenience and snowball sampling. This approach provided us with a potentially limited sample of participants. Hence, our sample might impact how our results generalize to other users. Future work should expand the sample to include different populations, especially underrepresented user groups, such as people with limited visual capabilities or difficulties using conventional keyboards (e.g., upper extremity impairment). Additionally, users of other smart home assistant systems, such as Google Home, might be worthy of further investigation.

We designed our scenarios around a subset of security-sensitive tasks on smart home assistants, namely online banking and smart door locks. As these tasks are currently available in certain markets, we hoped participants might have made some experiences with them. Abdi et al. [3] identified additional security-sensitive tasks, which future work should investigate upon, considering the different circumstances users might experience during the interaction. The most noteworthy tasks, that we did not investigate in our study, are healthcare and home surveillance, as users perceived them as most sensitive.

Some of our scenarios, such as the *"Dinner"* scenario, might not depict real-world use cases that users would want to engage in of their own accord. We deliberately chose edge cases for our study to provoke a stronger reaction from the participants and get richer data. Some of our scenarios include 4-digit PIN authentication, as this is the current standard method on the Alexa platform. However, as other voice assistants include different default settings (see Abdi et al. [2]), future work might benefit from investigating how these different authentication settings impact users' perceptions.

We cope with potential bias introduced by our personal expectations by making our them explicit in Section 3.5.

6 Conclusion

Our interviews explored the design space of authentication for smart home voice assistants. As security-sensitive tasks gain traction on this platform, developers and users call for appropriate authentication measures that enable privacy-preserving functionality and protect data from unauthorized access. Currently used authentication methods such as voice codes and biometric voice recognition proved insufficient considering both casual and targeted attackers. Prior work has already proposed some authentication schemes. However, no previous work has investigated the requirements for authentication systems from a users' perspective.

We closed this gap in the literature by reporting the results of a qualitative user study focusing on security-sensitive tasks on Amazon Alexa. We conducted 16 semi-structured interviews that included four scenarios involving high-risk tasks with Alexa users about (1) their perceptions of threats, (2) mitigation strategies, and (3) design factors that impact secure interaction experience. By performing a thematic analysis, we found that users are primarily concerned about bystanders that can eavesdrop on their interaction with Alexa. Our participants strongly favored biometric voice recognition as they perceived it as a natural and unobtrusive form of authentication. However, most users noted that current systems were not satisfying their security requirements due to being vulnerable to familiar attacks such as the *Replay Attack*.

Based on the insights gained from our user study, we provided design recommendations for future authentication systems. One such recommendation is based on a key finding that users have context-dependent requirements for authentication on smart home assistants. Users perceived levels of risk depending on the location of the interaction (e.g., inside the home vs. outside) and the type of bystanders (e.g., family members vs. casual acquaintances). Participants valued effortless and straightforward interaction with smart home voice assistants. Hence, authentication methods should strictly avoid distracting from primary tasks.

As this study is exploratory, future work can evaluate the findings on a broader basis. Users who have difficulties using traditional computing devices, such as users who can not read well, or users with visual impairment, rely on smart home voice assistants for their daily computing needs. The security and privacy needs and perceptions of this understudied group should be considered in future work.

Acknowledgments

We thank our study participants, as well as our interview partners for the pilot study. Thanks to Simon Anell for helping with transcribing the interviews and Florian Fankhauser for providing feedback on an earlier version of this work. Lastly, we thank the anonymous reviewers and our anonymous shepherd for their valuable and constructive feedback, which was very useful in improving our paper.

References

[1] ATI Scale. https://ati-scale.org/. [Accessed: 2021-02-25].

[2] Noura Abdi, Kopo M. Ramokapane, and Jose M. Such. More than Smart Speakers: Security and Privacy Perceptions of Smart Home Personal Assistants. In *USENIX Symposium on Usable Privacy and Security (SOUPS) 2019*, SOUPS, pages 451–466, Santa Clara, CA, USA, 2019. USENIX Association.

[3] 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*, CHI, pages 1–14, Yokohama, Japan, 2021. ACM.

[4] Ruba Abu-Salma, Elissa M. Redmiles, Blase Ur, and Miranda Wei. Exploring User Mental Models of End-to-End Encrypted Communication Tools. In *8th USENIX Workshop on Free and Open Communications on the Internet (FOCI 18)*, FOCI, Baltimore, MD, USA, 2018. USENIX Association.

[5] Amazon. Add Personalization to Your Skill | Alexa Skills Kit. https://developer.amazon.com/de/docs/custom-skills/add-personalization-to-your-skill.html. [Accessed: 2021-02-25].

[6] Simon Anell, Lea Gröber, and Katharina Krombholz. End User and Expert Perceptions of Threats and Potential Countermeasures. In *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, EuroUSEC, Genoa, Italy, September 2020. IEEE.

[7] August. Control Your August Smart Lock with Amazon Alexa | August. https://august.com/pages/alexa. [Accessed: 2021-02-25].

[8] 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)*, FOCI. USENIX Association, 2020.

[9] Logan Blue, Hadi Abdullah, Luis Vargas, and Patrick Traynor. 2MA: Verifying Voice Commands via Two Microphone Authentication. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, AsiaCCS, pages 89—100, Incheon, Korea, 2018. ACM.

[10] Anthony J. Bradley. Brace Yourself for an Explosion of Virtual Assistants. https://blogs.gartner.com/anthony_bradley/2020/08/10/brace-yourself-for-an-explosion-of-virtual-assistants/, August 2020. [Accessed: 2021-02-25].

[11] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006.

[12] Capital One. Capital One is on Amazon Echo. Questions? Just ask Alexa. https://www.capitalone.com/applications/alexa/. [Accessed: 2021-02-25].

[13] George Chalhoub and Ivan Flechais. "Alexa, Are You Spying on Me?": Exploring the Effect of User Experience on the Security and Privacy of Smart Speaker Users. In *HCI for Cybersecurity, Privacy and Trust*, HCII, pages 305–325, Copenhagen, Denmark, 2020. Springer International Publishing.

[14] Huan Feng, Kassem Fawaz, and Kang G. Shin. Continuous Authentication for Voice Assistants. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*, MobiCom, pages 343—355, Snowbird, UT, USA, 2017. ACM.

[15] 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.

[16] Marian Harbach, Sascha Fahl, and Matthew Smith. Who's Afraid of Which Bad Wolf? A Survey of IT Security Risk Awareness. In *2014 IEEE 27th Computer Security Foundations Symposium*, pages 97–110, Vienna, Austria, 2014. IEEE.

[17] David Harborth and Sebastian Pape. German Translation of the Concerns for Information Privacy (CFIP) Construct. *SSRN Scholarly Paper*, (ID 3112207), 2018.

[18] Yue Huang, Borke Obada-Obieh, and Konstantin (Kosta) 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*, CHI, pages 1–13, Honolulu, HI, USA, 2020. ACM.

[19] 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 *2019 IEEE Symposium on Security and Privacy (SP)*, SP, pages 246–263, San Francisco, CA, USA., 2019. IEEE.

[20] Il-Youp Kwak, Jun H. Huh, Seung T. Han, Iljoo Kim, and Jiwon Yoon. Voice Presentation Attack Detection through Text-Converted Voice Command Analysis. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI, pages 1–12, Glasgow, UK, 2019. ACM.

[21] 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):102:1–102:31, 2018.

[22] Xinyu Lei, Guan-Hua Tu, Alex X. Liu, Chi-Yu Li, and Tian Xie. The Insecurity of Home Digital Voice Assistants - Vulnerabilities, Attacks and Countermeasures. In *2018 IEEE Conference on Communications and Network Security (CNS)*, CNS, Bejing, China, May 2018. IEEE.

[23] Karen Levy and Bruce Schneier. Privacy threats in intimate relationships. *Journal of Cybersecurity*, 6(1):1–13, 2020.

[24] Nathan Malkin, Joe Deatrick, Allen Tong, Primal Wijesekera, Serge Egelman, and David Wagner. Privacy Attitudes of Smart Speaker Users. *Proceedings on Privacy Enhancing Technologies*, 2019(4):250–271, 2019.

[25] Diogo Marques, Ildar Muslukhov, Tiago Guerreiro, Luís Carriço, and Konstantin Beznosov. Snooping on Mobile Phones: Prevalence and Trends. In *12th Symposium on Usable Privacy and Security (SOUPS 2016)*, SOUPS, pages 159–174, Denver, CO, USA, 2016. USENIX Association.

[26] Dennis Reineck, Volker Lilienthal, Annika Sehl, and Stephan Weichert. Das faktorielle Survey. Methodische Grundsätze, Anwendungen und Perspektiven einer innovativen Methode für die Kommunikationswissenschaft. *M&K Medien & Kommunikationswissenschaft*, 65(1):101–116, 2017.

[27] Joshua Reynolds, Trevor Smith, Ken Reese, Luke Dickinson, Scott Ruoti, and Kent Seamons. A Tale of Two Studies: The Best and Worst of YubiKey Usability. In *2018 IEEE Symposium on Security and Privacy (SP)*, SP, pages 872–888, Oakland, CA, USA, 2018. IEEE.

[28] Ellen A. Rose. An examination of the concern for information privacy in the New Zealand regulatory context. *Information & Management*, 43(3):322–335, 2006.

[29] Nirupam Roy, Sheng Shen, Haitham Hassanieh, and Romit R. Choudhury. Inaudible Voice Commands: The Long-Range Attack and Defense. In *15th USENIX Symposium on Networked Systems Design and Implementation (NSDI 18)*, NSDI, pages 547–560, Renton, WA, USA, 2018. USENIX Association.

[30] 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 *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*, SOUPS, pages 211–228, Santa Clara, CA, USA, 2017. USENIX Association.

[31] Lea Schönherr, Maximilian Golla, Thorsten Eisenhofer, Jan Wiele, Dorothea Kolossa, and Thorsten Holz. Unacceptable, where is my privacy? Exploring accidental triggers of smart speakers. *arXiv preprint arXiv:2008.00508 [cs.CR]*, 2020.

[32] Jeff H. Smith, Sandra J. Milberg, and Sandra J. Burke. Information Privacy: Measuring Individuals' Concerns About Organizational Practices. *MIS Q.*, 1996.

[33] Elizabeth Stobert and Robert Biddle. The Password Life Cycle: User Behaviour in Managing Passwords. In *10th Symposium On Usable Privacy and Security (SOUPS 2014)*, SOUPS, pages 243–255, Menlo Park, CA, USA, 2014. USENIX Association.

[34] Takeshi Sugawara, Benjamin Cyr, Sara Rampazzi, Daniel Genkin, and Kevin Fu. Light commands: Laser-Based audio injection attacks on voice-controllable systems. In *Proceedings of the 29th USENIX Security Symposium*, SEC, pages 2631–2648. USENIX Association, 2020.

[35] Shari Trewin, Cal Swart, Larry Koved, Jacquelyn Martino, Kapil Singh, and Shay Ben-David. Biometric authentication on a mobile device: A study of user effort, error and task disruption. In *Proceedings of the 28th Annual Computer Security Applications Conference*, ACSAC, pages 159–168, Orlando, FL, USA, 2012. ACM.

[36] Rick Wash. Folk models of home computer security. In *Symposium on Usable Privacy and Security (SOUPS)*, SOUPS, pages 1–16, Redmond, WA, USA, 2010. USENIX Association.

[37] 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, CHI, pages 1–12, Glasgow, UK, 2019. ACM.

[38] 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):59:1–59:24, 2019.

[39] Eric Zeng, Shrirang Mare, and Franziska Roesner. End User Security and Privacy Concerns with Smart Homes. In *13th Symposium on Usable Privacy and Security (SOUPS 2017)*, SOUPS, pages 65–80, Santa Clara, CA, USA, 2017. USENIX Association.

[40] Eric Zeng and Franziska Rösner. Understanding and Improving Security and Privacy in Multi-User Smart Homes: A Design Exploration and In-Home User Study. In *Proceedings of the 28th USENIX Security Symposium*, SEC, pages 159–176, Santa Clara, CA, USA, 2019. USENIX Association.

[41] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. DolphinAttack: Inaudible Voice Commands. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS, pages 103—117, Dallas, TX, USA, 2017. ACM.

[42] Linghan Zhang, Sheng Tan, and Jie Yang. Hearing Your Voice is Not Enough: An Articulatory Gesture Based Liveness Detection for Voice Authentication. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS, pages 57—71, Dallas, TX, USA, 2017. ACM.

[43] Verena Zimmermann, Merve Bennighof, Miriam Edel, Oliver Hofmann, Judith Jung, and Melina von Wick. 'Home, smart home' – exploring end users' mental models of smart homes. In *Mensch Und Computer 2018 - Workshopband*, MuC, Dresden, Germany, 2018. Gesellschaft für Informatik e.V.

A Interview Guideline

The guideline we used for our interviews looked as follows, note that italic text indicates actions taken by the interviewer.

Introduction

Greet participant and introduce topic: "Hi, thank you for taking part in this interview." *Present interviewee with consent sheet, explaining purpose of the study.* "In the following, I will ask you some questions where I'm interested in your personal opinions and experiences, so keep in mind there are no wrong answers. If you feel like drawing anything throughout the interview, feel free to use this pen and paper here. Do you have any questions?" *Answer questions of interviewee, if any.* "So let's start with the first question!"

- How long are you using Alexa already?
  - **Alternative:** When was your first contact with Alexa?
- What devices are you using Alexa on?
- Where are those devices usually located?
- What are some typical tasks you perform with Alexa?
- Did you ever use Alexa for online shopping?
  - **If yes:** Did you encounter any issues while doing so?
  - **If no:** Where there specific reasons for you not to use this feature?

Scenarios

Lead over to scenarios: "Thank you for your answers so far. Now I would like you to have a look at some scenarios. For this interview, let's assume that all of the following features are implemented in Alexa, even though some of them are not currently available."

"Now I would like you to please take one of the scenario cards, have a look at it and read it aloud." *Let interviewee choose a card and flip it over.*

For each scenario:

- Please identify any issues that could arise in such a situation?
  - **Follow up:** Why do you think that is problematic?
- Can you identify threats for the user in such a scenario?
- Who could be the source of such a threat?
- What would you do to protect yourself?

Transition to next scenario: "Great, let's continue with the next scenario. However, we can always come back to a previous scenario if you want to add something." *Repeat process for all four scenarios.*

After all four scenarios:

- Now that you have seen all four scenarios, what do you think they have in common?

Demographics

Conclude scenario part and retrieve demographic data: "Thank you for the collaboration so far. Please take the tablet and fill out the questionnaire there." *Hand tablet to interviewee to complete the questionnaire.*

- ATI scale [1]
- CFIP scale [17]
- How old are you? *[free response]*
- What is your gender? *[free response]*
- What is the highest education you have completed? *[Single-select]*
  - *Elementary school*
  - *Junior High school*
  - *High school*
  - *Bachelor's degree or equivalent*
  - *Master's degree or equivalent*
  - *PhD*
  - *Other: [free response]*
- How many people live in your household? *[free response]*
- How many of them use Alexa? *[free response]*

Debriefing

- Do you have any final questions or marks you would like to make?

Thank interviewee for their collaboration and bid farewell: "Thank you again for your participation and have a nice day!"

B Codebook

The following list shows all codes and their categories we used for analysis. The brackets next to the code signify the overall number of occurrences in the interviews.

- **Attackers and Threats**
  - Accidents as threat (49)
  - Amazon listening in on conversations (8)
  - Bystanders as threat (51)
  - Criminals as threat (42)
  - Cyberattacks as threat (40)
  - Insiders as threat (39)
  - Malicious skills as threat (2)
  - Pranks as threat (19)
  - Sharing data with Amazon undesirable (95)
- **Biometric Authentication**
  - Annoyance of false negatives when using biometrics (3)
  - Authentication via voice recognition desirable (50)
  - Risk of false positives when using biometrics (16)
  - Uncertainty about security of voice recognition (16)
  - User wants Alexa in combination with face recognition (17)
  - Voice recognition should distinguish live voice from replays (4)
- **Building Trust**
  - Build/Lose trust through interaction experience (51)
  - Build trust in security mechanism via trial-and-error (6)
  - Trust from reviews (5)
  - Trust in familiar people (41)
  - Trust in system is transferred from trustworthy entity (39)
- **Knowledge-based Authentication**
  - Enter voice code via smartphone rather than Alexa (25)
  - High number of voice codes difficult to remember and distinguish (30)
  - User wishes for duress code (2)
  - Voice code protects against unauthorized access (27)
  - Whispering the voice code protects against eavesdropping (3)
- **Optimistic Authentication**
  - Optimistic authentication does not protect from physical access (1)
  - Optimistic authentication via delayed verification (35)
- **Perceptions of Alexa**
  - Insufficient mental model (17)
  - Personification of Alexa (15)
  - Uncertainty about security of Alexa ecosystem (22)
- **Perceptions of Authentication**

  - Properties of authentication method are transferred from other systems (86)

- **Possessions-based Authentication**
  - Risk of Replay Attacks when using tokens (2)
  - User wishes for Alexa in combination with OTP (25)
  - User wishes for Alexa in combination with token-based authentication (39)

- **Public Sphere of Alexa Interaction**
  - Openness of voice interaction security/privacy relevant (85)
  - Reconnaissance of Alexa easily possible (4)

- **Requirements of Authentication**
  - Multiple users use Alexa in parallel (10)

- **Risk Assessment of Alexa Authentication**
  - Minimal protection by law (8)
  - Users notice acoustic attacks on their Alexa if they are present (3)
  - User wishes for multiple authentication steps (37)
  - Variable security requirements depending on location (42)
  - Variable security requirements depending on presence of user (9)
  - Weighing up risks and effort of authentication (65)

- **Risk-Benefit Analysis of Alexa**
  - Alexa needs justification to exist (117)
  - Refrain from using the system due to security reasons (37)
  - Weighing up use against increased exposure to risk (30)

- **Social Aspects of Alexa Use**
  - Hierarchy among Alexa users (1)
  - Take time for important actions (6)
  - Using Alexa means being lazy (24)
  - Voice interaction inappropriate in specific social situations (31)

- **Transparency and Agency**
  - User wishes for agency over transparent processes (86)

- **Users' Mitigation Strategies**
  - Build trust in security mechanism via trial-and-error (6)
  - Change voice code regularly (11)
  - Move to another room to use Alexa (30)
  - Refrain from using the system due to security reasons (37)
  - Take time for important actions (6)
  - Users notice acoustic attacks on their Alexa if they are present (3)
  - Voice code protects against unauthorized access (27)
  - Voice interaction inappropriate in specific social situations (31)
  - Whispering the voice code protects against eavesdropping (3)

C Scenarios

Figure 1 shows the scenarios used in all our semi-structured interviews. We printed these for in-person meetings, we showed them on a website for online meetings, and made an audio version that included image descriptions for a blind participant.

You gathered some friends for a dinner party at your place. In the middle of eating you remember that you owe Kim 20€ for the lunch she paid the other day. You want to settle this right away.
You say: *„Alexa, transfer 20€ to Kim!"*
Alexa responds with: *„OK, to transfer money, tell me your voice code!"*
You: *„My code is 8915."*
Alexa accepts the code and the transaction succeeds.

(a) Dinner

You are in your living room watching TV when your partner asks, if you have already paid the utility bill this month. Since you have in fact not done so yet, you decided to do it right away using your Alexa device.
You say: *"Alexa, pay the utility bill!"*
Alexa answers: *"OK, to pay it, tell me your code!"*
You: *"6858"*
Alexa accepts the code and the payment is processed.

(b) TV

You have just taken all your groceries out of the car and are about to take them inside. The front door is locked. Your hands are full and you don't want to put everything down again so you ask Alexa to do open it for you.
You say: *"Alexa, unlock the front door!"*
Alexa answers: *"OK, to unlock the door, tell me your voice code!"*
You: *"3071"*
Alexa confirms the code and the door is unlocked.

(c) Door

You just came back from working in the garden. Your kids run around the house screaming. They are already very excited for the upcoming school trip. That's when the question comes to your mind: have you already paid for that? You want to check if the transaction is there in your online-banking.

(d) Hands

Figure 1: Scenarios used in the semi-structured interview