Understanding the Security and Privacy Implications of Online Toxic Content on Refugees

Arjun Arunasalam, *Purdue University;* Habiba Farrukh, *University of California, Irvine;* Eliz Tekcan and Z. Berkay Celik, *Purdue University*

https://www.usenix.org/conference/usenixsecurity24/presentation/arunasalam

This paper is included in the Proceedings of the 33rd USENIX Security Symposium.

August 14–16, 2024 • Philadelphia, PA, USA

978-1-939133-44-1

Open access to the Proceedings of the 33rd USENIX Security Symposium is sponsored by USENIX.

Understanding the Security and Privacy Implications of Online Toxic Content on Refugees

Arjun Arunasalam[†*], Habiba Farrukh[‡*], Eliz Tekcan[†*], and Z. Berkay Celik[†]
[†] *Purdue University, {aarunasa, etekcan, zcelik}@purdue.edu*
[‡] *University of California, Irvine, habibaf@uci.edu*

Abstract

Deteriorating conditions in regions facing social and political turmoil have resulted in the displacement of huge populations known as refugees. Technologies such as social media have helped refugees adapt to challenges in their new homes. While prior works have investigated refugees' computer security and privacy (S&P) concerns, refugees' increasing exposure to toxic content and its implications have remained largely unexplored. In this paper, we answer how toxic content can influence refugees' S&P actions, goals, and barriers, and how their experiences shape these factors. Through semi-structured interviews with refugee liaisons (n=12), focus groups (n=9, 27 participants), and an online survey (n=29) with refugees, we discover unique attack contexts (e.g., participants are targeted after responding to posts directed against refugees) and how intersecting identities (e.g., LGBTQ+, women) exacerbate attacks. In response to attacks, refugees take immediate actions (e.g., selective blocking) or long-term behavioral shifts (e.g., ensuring uploaded photos are void of landmarks) These measures minimize vulnerability and discourage attacks, among other goals, while participants acknowledge barriers to measures (e.g., anonymity impedes family reunification). Our findings highlight lessons in better equipping refugees to manage toxic content attacks.

1 Introduction

Deteriorating socio-political conditions in many countries have resulted in a crisis of displacement. Many victimized populations have been forced out of their home countries, seeking shelter and safety in neighboring or distant regions. Referred to as refugees, this population has fled unsafe conditions in search of a new home. More recently, the usurpation of power in Afghanistan and the invasion of Ukraine have led to a refugee crisis impacting over 11 million refugees [4, 85].

[*] Authors Arunasalam, Farrukh and Tekcan have made equal contributions to this work.

Refugees share a commonality in their experiences as a vulnerable population. They face challenges in displacement (e.g., adjusting to a new culture and language, economic disadvantage, and power dynamics due to legal status). Although members within this population have varying levels of digital literacy, prior work has shed light on how technology use amongst refugees benefits them [7, 53, 94]. As such, it has become evident that refugees widely use technology and, by extension, social media. For instance, work has highlighted how refugees use social media to organize their migration and build a life in their new homes [6, 40, 55].

Forced displacement of refugees has propelled this population to the center of many conversations on social media, which can be neutral or positive in nature. However, toxic content about/against refugees has recently proliferated, fueled by social, economic, and political factors [11, 87] and fake news [69]. Attackers emerge in online spaces, perpetrating toxic content to express hate against refugees, intimidate, and even bully individual refugees, among other reasons.

Toxic content attacks are a form of online hate and harassment [81], and can negatively affect individuals (e.g., heightened anxiety, depression) [77, 78]. However, the experiences of marginalized populations, i.e., refugees, can lead to intricacies in toxic content exposure and security and privacy-driven responses. Prior research has focused on online hate's impact on general populations and groups such as content creators [81, 82], while another line of work has explored understanding narrative justifications behind toxic content against refugees [11, 87]. Within the context of refugees' computer security, one work examined challenges the population faces, e.g., low literacy, to protect their computer security and privacy as newly resettled refugees [73]. However, research examining how refugees' experiences can produce nuanced toxic content exposure and how this exposure influences their digital behavior, i.e., responses and barriers that prevent security-privacy mechanisms, is largely absent.

In this paper, we conduct the first study to explore: **what are the impacts of toxic content exposure on refugees and corresponding security and privacy measures they take?**

To answer this, we first conducted semi-structured interviews with refugee liaisons (n=12) - individuals who work closely with refugees from four countries. These interviews provided insight into how toxic content affects refugees' digital presence and their responses. We then conducted focus groups (n=9, 27 participants) and an online survey (n=29) with refugees themselves to gain a deeper understanding of the specific goals of refugees' actions and the barriers they face. Our recruitment faced challenges due to the load interested parties faced caused by the recent and tragic Afghanistan and Ukrainian refugee crises. However, over one year, we interviewed/surveyed diverse refugee liaisons and refugees whose experiences were grounded in their use (or served refugees' use) of social media, exposure, and response to toxic content.

Our qualitative coding and quantitative analysis revealed novel insights into refugees' experiences on social media. Participants' social media dependency extends beyond general population use; they use it to source aid from support groups (other refugees and NGOs), and to rejoin family/friends separated through the refugee crisis. Participants' toxic content exposure coincides with the emergence of online conversations centered on the refugee crisis; they find themselves targeted after commenting on hate directed at refugees by attackers, who infiltrate closed groups. Intersecting identities (e.g., gender, sexuality) exacerbate threats, which are often perpetrated by strangers but can also originate from affiliates (e.g., neighbors) due to ambiguity behind one's refugee status.

Participants take security and privacy responses towards attacks, broadly categorized into immediate responses, e.g., selective blocking of attackers, using various reporting channels, and long-term behavioral changes, e.g., platform withdrawal and rigorous privacy measures. Participants' experiences result in response intricacies. For instance, participants express how toxic content against refugees is interpreted/disguised as expressing a political opinion, resulting in unfavorable results after reporting. Similarly, decisions to ignore toxic content are due to power imbalance; participants fear responding against these attackers, who are often residents/citizens.

Our participants also express goals to minimize vulnerability and discourage toxic content while also detailing barriers associated with privacy and anonymity (e.g., inability to change usernames as it complicates family reunification). Participants also advocate for platform involvement in mitigating toxic attacks while also acknowledging shortcomings in existing methods. They raise themes such as flawed censorship due to the contextual nature of toxicity while expressing distrust of automated toxic content detection on online platforms.

Our study extends efforts to understand at-risk users' experiences and how they shape S&P concerns. We conclude by synthesizing lessons on toxic content threat models, detection, and enforcement for online platforms to create a better online experience for refugees. We outline the necessity for refugee aid organizations to disseminate security and privacy advice on toxic content for refugees.

In this paper, we make the following contributions:

- We design protocol to understand toxic content's S&P implications on refugees. We conduct 12 semi-structured interviews with refugee liaisons and leverage insights to organize nine focus groups with 27 refugees and survey 29 refugees.
- We understand how refugees' experiences shape their (1) toxic content threats, (2) corresponding responses and goals, (3) barriers, and (4) mitigation perspectives.
- We synthesize lessons for (1) threat models, (2) toxic content detection and mitigation, and (3) S&P guidelines. We also outline future directions and interventions to aid platforms and NGOs in combating toxicity.

2 Background

2.1 Refugees and Liaisons

In this paper, we use the UN Refugee Agency's (UNHCR) definition of refugee, "someone who is unable or unwilling to return to their country of origin owing to a well-founded fear of being persecuted for reasons of race, religion, nationality, membership of a particular social group, or political opinion" [86]. The country a refugee flees to is referred to as the "hosting country". As of 2022, there are 35.3 million refugees, comprising many nationalities, cultures and religions [66].

Refugee liaisons are those with established close working relationships with refugees, either through professional/volunteer service or interaction with them. They encompass various professions, e.g., medical doctors and lawyers who predominantly serve refugees. Refugee liaisons are also NGO workers who provide refugees aid/services, and researchers such as academics who closely interact with refugees for academic study. For our study, we only consider liaisons that: (1) have had direct interaction with refugees through service, research or work, and (2) have observed and discussed social media use with refugees.

2.2 Toxic Content

Toxic content is an online attack, that falls under online hate and harassment. Thomas et al. define toxic content as "a wide range of attacks involving media (e.g., images, text) that attackers send to a target or audience, without the necessity of more advanced capabilities (e.g., does not require privileged access or deception)" [81]. Toxic content has the ability to impose a variety of harm, such as damaging one's reputation, affecting sexual safety, coercion, and intent to silence. Toxic content can also be broken down into different types of attacks (e.g., bullying, hate speech, and threats of violence). For instance, trolling is defined as toxic content that intentionally

provokes someone/group of people with inflammatory remarks. A complete description of toxic content subcategories we cover in our paper can be found in Appendix Table 1.

The ability of toxic content to spread on online platforms has incentivized guidelines that discourage toxic content posts. For instance, social media platforms such as Twitter have hateful conduct guidelines that disallow *"violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation ... [other at-risk minorities]"* [46]. Despite these guidelines, toxic content still proliferates in online spaces, harming various groups [43, 59], including refugees. Toxic content targeting refugees is defined as toxic content explicitly directed at a refugee or a group of refugees. In this paper, we refer to instances of such toxic content as "toxic content attacks", or just "attacks".

3 Related Work

Toxic Content and Online Hate. Researchers have extensively studied toxic content and other forms of online hate and harassment [81]. Several efforts have characterized and measured hate on platforms (e.g., Twitter, 4chan) [16, 18, 48, 65], and consider how factors such as attacker anonymity influence attacker behavior [96]. Hateful sentiment against specific groups such as women, the Jewish community, and other minorities [21, 88, 95] has also been studied. Similarly, Thomas et al. analyzed how content creators (online figures who produce media) deal with hate and harassment [82]. Another line of work has studied toxic content posts against refugees on online platforms, focusing on narrative justifications for anti-refugee sentiment [11, 52, 63, 83, 87]. However, the scope of their findings is limited to specific geographical regions and does not consider the effect this content has on its target - refugees. Community efforts have also explored automated mechanisms to detect toxic content [20, 26, 30, 32, 60] using machine learning-driven approaches.

Our work differs as prior efforts do not consider when targets of toxic content are refugees. Our research questions (which are previously unexplored) focus on how refugees' experiences shape the way they handle toxic content attacks and can differ from other groups (e.g., general users). Our protocol is also designed to gain insight into toxic content targeting refugees specifically because of their refugee status - we do not consider unrelated content (e.g., attacks solely exhibiting xenophobia/racism). We also understand refugees' perspectives on how toxic content mitigation strategies fail to meet the needs of this population.

At-Risk Users. The security community has long paid attention to at-risk users - individuals with "risk factors that augment their chances of being digitally attacked" [90]. Prior work has explored S&P factors and digital safety amongst communities including (but not limited to) domestic migrant workers [74], undocumented immigrants [42], LGBTQ+ communities [39, 72] and survivors of sexual assault [8, 61]. In understanding these communities, prior work has also shed light on the digital threats they face and how they navigate digital spaces (e.g., what data is considered sensitive) [15, 19, 36, 47]. Community responses to threats (e.g., blocking, social pleas), accompanying barriers [29, 57, 90], and how these individuals represent their digital selves (e.g., how the LGBTQ+ community identifies themselves online [34, 91]) have also been studied. Similarly, Simko et al. [73] investigated S&P for refugees and their reliance on digital technology (e.g., emails), interaction with threats (e.g., scams), and security practices (e.g., password creation techniques).

These efforts leverage interviews/focus groups to understand community-specific risks affecting marginalized populations and resulting threat models. We extend by focusing on (a) refugees, and (b) characterizing a highly specialized threat of toxic content. Our findings show how refugees have different experiences, S&P actions, and priorities than other at-risk communities when interacting with toxic content.

Refugees' Technology Usage. Community efforts to explore refugee interaction with technology have focused on how refugees use technology within refugee camps [93], and how refugees benefit from technology (e.g., smartphones, communication apps) in regard to the refugee crisis [40, 41, 80]. Efforts to outline challenges in HCI-refugee relations have also been conducted (e.g., challenges of computer club initiatives in refugee camps) [1, 2]. There also exists a body of work analyzing specialized use cases. For instance, Dyden-Peterson et al. [31] and Dahya et al. [25] focused on how mobile phones serve as an educational aid in refugee camps. Prior work has also examined refugee social media use in decision-making as they experience displacement [27, 40], and in grassroots/advocacy efforts [28].

We contrast prior research as we study online toxic content S&P implications on refugees' use of social media. Our carefully designed protocol understands how refugees use technology to enact S&P-focused actions and what S&P barriers exist in the technology they use.

4 Motivation and Research Goals

Refugees face unique challenges (e.g., fleeing from war, lacking an understanding of the local language, and economic hardship). These vulnerabilities are compounded by recent events (e.g., Afghanistan and Ukrainian refugee crises) propelling toxic rhetoric targeting them into mainstream channels (e.g., social media, news). We aim to unpack toxic content targeting refugees and answer the following research question:

RQ What are the impacts of exposure to toxic content on refugees and the corresponding security and privacy measures refugees take in response?

Given that we aimed to achieve an in-depth exploration of this research question, we divided it into four sub-questions:

SQ1 How do refugees interact with online toxic content and what actions do they take upon exposure?

SQ2 What are the security and privacy goals of refugees regarding toxic content?

SQ3 What security practices are barriers for refugees?

SQ4 What mitigation efforts would help refugees feel safer?

To answer these questions, we leverage a mixed-methods approach to investigate toxic content and its impact on this population. Through semi-structured interviews with n=12 refugee liaisons from four countries, we understand the security measures refugees take after exposure to toxic content. We then conducted focus groups with refugees (n=9, 27 participants) to validate liaisons' perspectives and obtain a detailed understanding (e.g., barriers in enacting countermeasures) across diverse participants.

Using these qualitative methods allows us to understand refugee experiences with toxic content from an S&P lens and the root causes of their decisions (e.g., reasons/goals behind actions) [38]. We also designed an online survey instrument (n=29) to provide an alternative for refugees who felt uncomfortable participating in focus groups. Answering our research question serves to help the S&P community better understand the threat of toxic content against refugees and what future directions researchers can take to combat this problem.

5 Study Methodology

We initially aimed to answer all four sub-research questions (**SQ1** - **SQ4**) by interviewing refugee liaisons, as they work with large and diverse refugee populations, and draw from years of experience. Their interactions provide us insight into how toxic content affects refugees' day-to-day lives in the digital world (**SQ1**) and refugee toxic content-related security and privacy goals (**SQ2**).

However, we find that liaisons are only able to provide generic answers to **SQ3** and **SQ4**. Refugees' toxic content-related S&P barriers are highly personal, and mitigation efforts that help refugees feel safer were hard to infer from their experiences working with refugees. To gain detailed insight into **SQ3** and **SQ4**, we followed liaisons' suggestions to conduct (1) focus groups with broad questions instead of topic guides and (2) online surveys for participants to maintain anonymity (available via *Qualtrics*).

5.1 Participant Recruitment

Refugee Liaison Interviews. We conducted semi-structured interviews with refugee liaisons, recruiting them from multiple countries and occupations, through a combination of snowball sampling and purposive sampling [68].

We reached out to 52 refugee liaisons from research foundations, refugee camps, and NGOs and also contacted personal contacts across eight countries: US, Turkey, Spain, Bulgaria, Germany, UK, France and Austria. 20 contacted liaisons did not respond. Among the 32 that did respond, 11 were unable to accommodate us due to scheduling difficulties. Nine liaisons communicated interest (via email) but believed they could not contribute due to lack of expertise in refugee social media use. The remaining 12 felt confident in their experience and observations regarding this sensitive topic.

Table 1: Overview of liaison participants.

#	Occupation	Experience (years)	Locality of Refugees Served†	Region of Service‡
P1	Lawyer	7	Middle East	Turkey
P2	Lawyer	8	Middle East	Turkey
P3	Medical Doctor	8	Middle East	Turkey
P4	Academic	41	Middle East	USA
P5	Academic	10	Middle East	Turkey
P6	Academic	5	Africa	USA
P7	NGO Worker	7	South America	Spain
P8	NGO Worker	14	Middle East	Bulgaria
P9	NGO Worker	2	Asia, Africa, Middle East	USA
P10	NGO Worker	4	Asia, Africa, Middle East	USA
P11	NGO Worker	3	Middle East	USA
P12	NGO Worker	7	Middle East	Turkey

† Region where refugees participants have worked with are from ‡ Region where liaison interacts with refugees (hosting country of refugees)

Table 1 overviews interviewed refugee liaisons. These liaisons comprise diverse professional experiences. NGO workers provide services to refugees via lessons (e.g., digital literacy, citizenship classes), coordinating operations in refugee shelters, or managing refugee aid (e.g., food, shelter). The lawyers we interviewed provided legal counsel to refugees in migration/resettlement centers. The liaison who worked as a doctor provided medical services in refugee camps. Academics conducted academic research studying refugee issues up close through workshops, fieldwork, and online interactions, offering unique insights compared to lawyers, NGO staff, and doctors. Interviewed refugee liaisons drew on many years of experience with an average experience of ~10 years and interactions with a large number of refugees.

Refugee Focus Groups. To recruit refugees, we contacted 21 NGOs that work with refugees through organizations and resettlement camps (across eight different countries). We also used our connections with all 12 interviewed liaisons. We prepared recruitment material, including a short video and a flyer, which were distributed to these parties. NGO workers and liaisons informed refugees of our study, with interested parties signing up to participate in our focus groups.

We faced challenges during the recruitment process. Due to the Afghan and Ukrainian refugee crises, NGOs had limited flexibility and time to coordinate and aid us in recruitment. However, over the course of one year, we were able to recruit a diverse participant pool for our focus groups. Table 2 describes the focus groups we conducted. We conducted nine focus groups (`FGs`) from three different affiliations. `FG1` comprised a family of three refugees. `FG2` comprised five refugees participating in government-sponsored lessons at a local university (where refugees learn the hosting country's native language). `FG3-9` consisted of 19 refugees participating in digital literacy/citizenship lessons conducted by an NGO.

Table 2: Overview of refugee focus groups conducted.

ID	Short Description	# Participants	# of Focus Groups	Region of Origin ‡	Hosting Country†
FG1	Family of settled refugees	3	1	Middle East	Turkey
FG2	Recently settled refugees at a language school	5	1	Middle East	Turkey
FG3-FG9	Settled refugees participating in NGO lessons	19	7	Asia (11), Middle East (5), South America (2), Africa (1),	USA

‡ Participants more commonly reported region of origin, instead of the country, when asked to self-report demographics † Country where participants currently reside

Table 3: Self-reported demographics of survey participants, reported as aggregates.

Region of Origins		Country of Settlement	Education Level	Gender Identity
Middle East 15	(Syria 8)	Turkey 17	Bachelor's degree 13	Male 15
	(Afghanistan 2)	USA 2	Associate degree 1	Female 11
	(Lebanon 3)	Dubai 1	High school degree or equivalent 4	n.a. 3
	(Turkey 1)	Netherlands 1	Less than a high school diploma 7	
	(Palestine 1)	Italy 1	Master's degree 1	
Europe 4	(Ukraine 3)	n.a. 7	n.a. 3	
	(Bulgaria 1)			
Asia 1	(Vietnam)			
South America 1	(Honduras)			
n.a. 8†				

† We mark n.a. (not applicable) for candidates who chose not to self-report specified demographics

Online Survey. We also provided the questions from our focus group in the form of an online survey. Survey recruitment was similar to that of focus groups. Our recruitment material included links to the online survey, as an alternative to focus groups. Liaisons in Table 1 distributed our survey. Here, we did not collect any organizational affiliation.

We prepared translated versions of our survey in Arabic, Turkish, Spanish, and Urdu. We chose these languages upon request from the 21 NGOs and 12 interviewed liaisons we coordinated with to recruit survey participants. Surveys were translated via Google Translate API [22], with translations verified/modified by recruited translators (fluent in the target language and English) referencing the English survey. Table 3 presents a detailed breakdown of the demographics of our participants. 29 refugees from the Middle East, Europe, Asia, and South America participated in our survey. Although an online survey limits direct interaction with refugees, it allowed us to consider perspectives from refugees who may not have felt comfortable sharing in a group setting.

5.2 Ethical Considerations

We worked with our IRB to ensure our materials were ethically designed. Our study was considered exempt after an initial review by a primary reviewer and analyst from the IRB office. Our study was designed after considering participants' privacy and the ethics of working with vulnerable populations.

First, we only collect personally identifiable information in recruitment (sign-up sheets) and our survey/interviews necessary for our study. For refugees, we collect self-reported age range, gender identity, education, and region of origin. For liaisons, we collected the region of origin of refugees they work with. Second, we de-identified transcripts by suppressing names and other personal data. Third, we audio-recorded interviews only after receiving the interviewees' consent. At every interview/focus group step, we ensured that dissemination of our data collection and results would not compromise the privacy and safety of participants, following heuristics to conduct research with vulnerable populations ethically [89].

We informed participants that they could skip any question due to the sensitive nature of our interview questions. Participants were not required to answer questions, and some chose the option of not answering. Translators acknowledged the unpleasantness of reading toxic content, but overall felt positive due to their research contribution.

5.3 User Study Procedures

Our questions focused on answering the four research questions we define (**SQ1**-**SQ4**). For both interviews and focus groups, our questions were purposefully wide and exploratory, allowing us to ask follow-up questions. For our online survey, we combined open-ended and multiple-choice questions.

Interviews with Liaisons. We conducted semi-structured interviews with liaisons remotely via video conferencing software, where each interview lasted 60 to 90 minutes. Each interview was conducted by two authors and audio recorded, transcribed and anonymized for data analysis. The interviews were conducted in English, Spanish, and Turkish. For non-English languages, we contacted researchers who were fluent in the language of the interviewee to join the interview, and they simultaneously interpreted our questions and interviewees' answers. We ask liaisons about their work with refugees, the types of refugees they work with, and their observations of toxic content directed at these refugees. A full list of initial questions for the interview is presented in Appendix B.1.

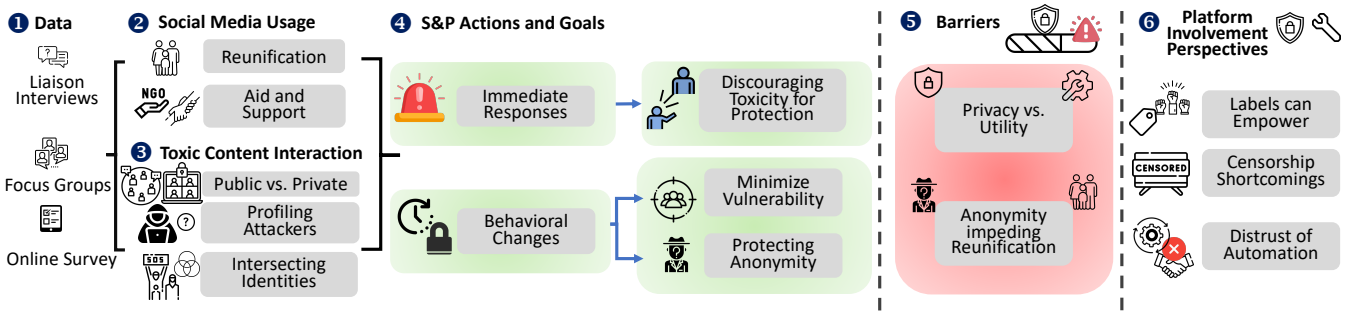

Figure 1: Overview of refugees' security and privacy implications due to toxic content, derived via a mixed-method approach. We discover nuances in participants' actions in response to toxicity, their goals, barriers, and perspectives on platform involvement.

Focus Groups with Refugees. Focus groups were conducted via video conferencing software and subsequently transcribed and anonymized. We did not allow third-party access (only study personnel) to transcripts. The interviews were conducted in English while working with native-speaker interpreters in the interviewees' language. We ask participants to self-report demographic data, and discuss their exposure to toxic content and responses. We also probed for their goals through these responses and factors preventing them from enacting responses. We note that data collection focused on content directed at participants because of their refugee status - we did not consider unrelated content (e.g., attacks *solely* exhibiting xenophobia and offline attacks). We conclude by asking for perspectives on platform mitigation. A full list of initial questions used can be found in Appendix B.2.

Online Survey. Our survey comprised two sections: (1) preliminary questions and (2) online toxic content-related questions asked in our focus groups. In (1), we ask participants to self-identify their status as a refugee to ensure only responses from refugees were considered.

5.4 Data Analysis

Transcripts. To thematically analyze transcripts (which were stored in a secure cloud service), we first used inductive coding. We produced one codebook each for liaisons and focus groups respectively. Two independent coders familiarized themselves with transcripts, and coded transcripts separately before discussing the code selections and settling on an intermediate codebook. Next, coders deductively coded using the online toxic content taxonomy (Appendix Table 1) and the intermediate codebooks to identify security measures refugees take. All interviews were double-coded, and coders stopped after every interview to discuss changes to the codes and themes until agreement. We do not present intercoder agreement as independently coded transcripts were reviewed together [58].

Through our intermediate codebook, we focused on (1) social media use, (2) toxic content exposure, (3) abusers and motivating sentiment for online hate and harassment, (4) toxic content impact and (5) refugees strategies for coping with online abuse. Coding occurred simultaneously with our recruitment and interviewing. We follow thematic saturation guidelines [71]. When no new codes emerged, and themes were repeated during coding, we ceased recruitment as additional interviews were unlikely to introduce new takeaways.

Survey Data. Non-English responses were translated by recruited translators fluent in response language. We code responses to open-ended questions while considering responses to multiple-choice questions. We merge themes extracted from open-ended questions with interviews/focus groups themes. When relevant to themes, we present percentages for multiple-choice questions.

6 Security and Privacy Impact on Refugees

Figure 1 presents an overview of our findings from interviews, focus groups, and online surveys (❶). We first outline participants' use of online platforms (❷). We then detail experiences with toxic content, focusing on toxic content types and contexts (e.g., where attacks occurred and what led to the incident) (❸). We also survey responses to toxic content, corresponding S&P goals (❹) and barriers (❺), and perspectives on mitigation efforts ❻). We detail below, referring to surveyed/interviewed refugees as *participants* and interviewed liaisons as *liaison participants*. For focus groups, we attribute findings to respective focus groups, not specific participants.

6.1 Online Platforms and Refugees

All participants we talked to had access to smartphones and/or a personal computer/laptop, which serve as a link to their online life. Devices are usually under the sole ownership or shared amongst families, as echoed in prior work [73]. For the purpose of our study, we consider social media as any online platform/app allowing users to interact with others, e.g., private one-on-one interactions, viewing publicly-shared content, and group settings (e.g., Telegram, WhatsApp). All participants partake in social media, using at least one social

media application daily. Facebook was the prime social media used (51% of survey participants and all participants in FG1-9). Social media platforms such as Twitter, Instagram, and TikTok were also popular. Participants also report using applications such as WhatsApp and Telegram to engage in social activities and communicate with friends/family. Social media was also used for a variety of secondary reasons (e.g., *"to find employment"*, *"for business"*). Refugee participants expressed confidence in using social media, suggesting digital literacy among our interview population. We note that we define digital literacy in the context of our research questions - we consider participants digitally literate if they self-reported comfort in using social media, and were aware of and understood how to access features (e.g., how to change privacy settings, how to block, how to report users).

Similar to general users, participants use social media for public activity, e.g., interacting/commenting on family/friends' posts or sharing a picture. Participants also engage in private settings, e.g., joining private groups. However, we find that, due to their experiences, participants' activity extends beyond common uses.

Means for Reunification. We find that participants rely on social media to stay connected with displaced family/friends in different regions, similar to social media use among immigrants [13, 42]. However, contrasting the experiences of other populations, refugee participants maintain a digital connection for future reunification - the rejoining of families separated by the refugee crisis. Here, participants note that the instability of the region they come from complicates this process (e.g., not always knowing where a loved one is currently residing) and thus increases dependency on social media. Participants express that interactions for the purpose of reunification usually begin as a public activity (e.g., commenting on a public post) but soon migrate to private settings.

Prior work highlights how refugees are dependent on technology, (e.g., mobile phones) to stay in touch with family/plan their journey [40, 41] and highlight limitations of using mobile phones to stay connected, due to dependency on SIM and mobile service providers [54]. We note that social media such as Facebook/Instagram are appropriate to aid in reunification as they overcome limitations (participants may not have access to a working phone number in all stages of their movement). The importance of staying online for reunification echoes other communities' needs to stay online for other vital needs, such as online creators who depend on online media for income [82] or trafficking victims who need social media to communicate with victim service providers [19].

Sourcing Aid and Support. Liaison participants highlighted how refugees prefer closed groups that only accept refugees/individuals affiliated with refugees. Liaisons participate in such groups to reach large groups of refugees and disseminate information. Participants use private groups as a source of support during resettlement - settling in another country as a refugee. Their involvement in these groups varies, ranging from requesting items for daily needs (e.g., house supplies) to accessing information on NGO-organized training programs or government aid such as stipend assistance.

6.2 Online Hate and Harassment

All participants acknowledged exposure to toxic content targeting the general refugee population, such as posts containing hateful sentiments targeting refugees. However, we discover that bullying (where the participant is targeted) is the more common attack they are exposed to. These targeted attacks usually begin as interactions with attacks targeting the general refugee population. For instance, participants note that commenting on a toxic post targeting refugees often leads to bullying. Interestingly, participants report that these attacks are often multi-fold, with a single post combining other secondary attacks, e.g., bullying *and* threats of violence.

Among secondary attacks, trolling and hate speech comprised the majority of toxic content participants experienced (FG1-9, 55% of survey participants). Attacks that leverage profane and offensive language were also common. Liaison participants noted racial slurs (offensive language) directed at the refugees they work with.

Other examples of attacks accompanying bullying include sexual insults and content mocking the participants' physical appearance. These attacks fall under sexual harassment and purposeful embarrassment. One liaison noted, *"The attackers use the word [race] as an insult; I see comments directed at my clients such as 'You are an ugly [race], you should go back to your country and many instances of similar comments"*

Liaison participants observe consistent interaction between refugees and toxic content; however, they also report that refugees face heightened exposure during specific seasons, e.g., an election season or religious holidays.

6.2.1 Public and Private Exposure to Toxic Content

Most participant interactions with toxic content occur in public settings, where any user with an online account can view this toxic content. 76% and 11.5% of survey participants experience toxic content attacks in public settings on social media and online forums, respectively.

Toxic content, however, also infiltrates private settings, where content can only be viewed by select users. 7.7% and 3.8% of survey participants faced toxic content on a dating website or messaging app, respectively. For instance, one participant using a dating app noted a benign experience that became toxic - *"There were people who harassed me [once finding out I] was a [refugee]. When [I refused] to meet ... [someone, they] insulted me."*.

Toxic content in private settings also affects participants within closed groups. Participants of one focus group reported an incident where their WhatsApp group (that can only be joined via invitation) had been *"hacked [infiltrated], and they*

started receiving offensive images targeting refugees." We posit this resulted from an unintentional leak of the group's "invite link", as speculated by participants. This experience of infiltration was the first of its kind for FG3, whose participants reported *"feeling very uneasy and worried about this attack."*

6.2.2 Profiling Attackers

In profiling attackers, participants highlight the following three factors: (*a*) affiliation with refugees, (*b*) association with organizations/ideology, and (*c*) demographics.

For affiliations, a majority of attacks against participants stem from strangers, people who have no prior affiliation with the refugee (e.g., FG1-9, 57% of surveyed participants had never met their attacker before). However, similar to victims of intimate partner violence [36], and trafficking [19], participants note more personal attackers. These include neighbors, acquaintances, or romantic interests. Participants expressed that the emergence of personal attackers stems from the ambiguity surrounding their refugee identity.

Prior work has studied self-disclosure of status in regards to online threats (e.g., some HIV-positive individuals are comfortable noting status in dating apps [91] while transgender individuals may opt against disclosure [34]). We find that participants opt against preemptive disclosure of their refugee status on their public profiles. Thus, perpetrators of attacks targeting the general refugee population may be unaware that the participant is a refugee and will be exposed to the attack. Participants expressed this can trigger an episode of bullying should the participant decide to respond. For instance, one participant who claimed that their neighbors were not aware of their status stated that *"neighbors made a [toxic] campaign against refugees [on] social media."* Association with political organizations was also a recurring theme when participants spoke of attackers. For example, one participant noted, *"[they] read a lot of posts contain[ing] hate speech about refugees on social media groups ... run by support[ers] of a [political party] in [country]."*

Liaisons specified that attackers' motivations vary based on their demographics, e.g., socioeconomic status. Upper-middle, middle, and working-class citizens in a hosting country have varying sentiments against refugees. One liaison highlighted how, *"If you were working class, you were more likely to worry about competition with [refugee] workers, [who are] willing to [or] exploited to work for less... but middle or upper middle class ... worry about invasion of cultural values."*

6.2.3 Intersecting Identities and Compounding Attacks

We observe that intersecting identities within refugee populations influence their toxic content experience, a theme consistent with the intersectionality framework [24, 51]. Participants with intersectional identities note that attacks are more severe as attackers leverage insults against additional identities beyond their refugee status.

Sexual Orientation and Gender. Sexual insults were commonly targeted at female participants, with one noting the violent threat of rape being a common threat (*" I have seen toxic comments like: 'We want to rape all the refugee women' "*). LGBTQ+ refugees face similar compounding threats. One liaison who interacted with refugees in Kenya noted the Free Block 13 campaign - an initiative to help persecuted LGBTQ+ refugees in a LGBTQ+-only camp, known as Block 13 Kakuma [35]. The campaign's social media presence on Facebook and involvement with LGBTQ+ refugees make it a prime target for attacks. Social media actions such as "tagging" another refugee result in targeted attacks against the tagged individual (an instance of bullying), suggesting that attacks vary between participants who are recently settled and those in refugee camps. In one event, a liaison noted *"[Someone] came on to the post about the evacuation of LGBTQ+ refugees to prevent persecution and attacked the refugee that was tagged in the post, basically saying, 'You know you don't belong here, you should leave, we don't want you here'."*

Language, Culture and Religion. Participants and liaison participants expressed how different cultures were prevalent themes in their toxic content experiences, with language and religion being common targets. Posting content in a language other than a country's native language can elicit toxic content responses, as attackers assume that the poster is a refugee. Liaisons noted that organizations are afraid to write social media posts in the *"language that refugees use"*, due to prior incidents where such posts were targeted by attackers.

Prior work [5] has highlighted how Muslim Americans are subjected to heightened religious discrimination. We extend these findings by showing that religion emerges as an important intersectional identity among the already marginalized group of refugees. For instance, one participant noted they *"posted something religious and several friends [wrote] negative things [about refugees]"*. Muslim participants noted insults against their religion often accompany attacks, suggesting demographic differences in what is considered toxic content (Christian participants did not consider attacks against their religion as toxicity against refugees).

The impact of intersecting identities is also exacerbated by recent phenomena. One liaison participant who primarily worked with Asian refugee students noted an uptick in complaints due to COVID-related anti-Asian sentiment. This sentiment has resulted in an increase in targeted attacks. This liaison expressed how *"[Their] Asian students were [reporting] having things said [directly] to them online."*

Despite a diverse participant pool, we acknowledge that our data might be limited to providing conclusive evidence of how intersectionality impacts refugee S&P implications and perceptions of toxic content (Detailed in Section 7.4). However, our data provide directional evidence of in-population differences (as outlined above), which warrants further study into how demographic differences can impact S&P implications.

6.3 S&P Actions and Goals

We discover a variety of responses to toxic content that participants take, grounded in preserving their security and privacy, as well as the goals associated with these approaches. Our findings can be divided into two types of actions: (*a*) immediate responses and (*b*) behavioral changes.

Immediate responses refer to the action a participant takes *immediately* upon seeing/interacting with toxic content. We find that participants' goals with a majority of immediate responses is to discourage toxic content. Behavioral changes refer to how toxic content leads to long-term changes in a participant's actions online, with their goal here being to either minimize vulnerability and/or protect their anonymity.

6.3.1 Immediate Responses to Attacks

We discover four immediate responses to toxic content attacks: (*a*) selective blocking, (*b*) engaging attackers, (*c*) using reporting channels, and (*d*) ignoring toxic content.

Selective Blocking. The most common immediate response by refugees who encounter toxic content online is to block the attacker (59% of surveyed participants). Participants state that they usually block attackers because they fear for their online safety (e.g., *"[I] got scared, and blocked [the attacker] after being threatened."*). While blocking is a common attack response mechanism for general populations and at-risk communities [29, 56, 70], and in some cases the *primary* form of privacy protection [67], this decision is seen as a "last resort" by participants. Participants prefer less critical immediate responses, such as "unfollowing" or removing themselves from the group. For instance, one participant noted that they would rather ignore the attacker than block as *"it is a softer resolution"*. Perception of an attack's severity can also influence a decision to block. Some participants noted that toxic content directed at refugees (instead of them alone) warrants a block. Other participants block attackers when bullied.

Engaging Attackers. Another immediate response that participants choose is to engage or communicate directly with the attacker (`FG3-9`). This response is similar to at-risk users' attempts at social pleas [90], but instead of requesting the attacker to stop attacks, participants expressed interest in engaging the attacker in conversation. Participants note that this is motivated by a desire to understand the attacker's motivation, e.g., *"I want to know why you are saying this"*, *"If you explain to me and I am wrong, I will say sorry."*. Contrasting work outlining preferences amongst online communities to avoid engagement [82], participants note that direct engagement is empowering, as it allows them to confront attackers who usually commit abuse unchallenged, e.g., *"if they just make fun or try to make [refugees] look bad, then I will [converse with] them on Facebook. That's the chance that I have."*

Direct engagement may escalate into further toxicity - participants express that attackers respond with further insults/profane language. One participant mentioned that their direct engagement attempt was met with efforts to "dox" them, with an attacker posting threats, *"You are from here [location]. This is your address? [address] I know your address"*. Despite this, the same participant also shared positive results from direct engagement, *"meeting [the attacker] in the middle after [the attacker] admitt[ed] [they] has been rude [and finally, they] said sorry."* Although attacker self-realization was uncommon (only 2 participants received apologies), accounts of these instances empowered other focus group participants who stated they were now more inclined to engage.

Reporting Channels. Participants' self-reported confidence in using social media extended to their familiarity in engaging with reporting channels (e.g., content or user profile reporting tools); 14% of surveyed participants stated they report attackers to the platform. For instance, one participant states *"Sometimes when [I] see this content, [I] will report it to Facebook or whatever the authorities of the application [I am] using"*. We find that participants depend on how a platform's policy is enforced, in the hopes that *"[using the feature] will result in the company [banning] that user."*.

Our findings complement research on user frustration with reporting [82] - participants express a lack of clarity on platform decisions when the reporting mechanisms they use yield no results. Participants attributed this to the ambiguity of what platforms consider toxic content. Participants believe that abuse directed at them is perceived as a contentious political discussion and exercise of free speech, especially in the United States, where censorship on social media and the right to express opinions are topics widely debated. To illustrate, one participant notes *"it's kind of tricky [to report] because I hear [America] has the First Amendment"*.

Unfavorable results (e.g., dismissed reports with the attacker facing no consequences) prompt participants to report to other authorities, such as law enforcement. This is especially true for more personal attacks and threats, including doxing and threats of violence. The involvement of law officers, however, causes further complications due to the legal status of the refugee, a hindrance not common among other communities [82]. For instance, one participant who experienced toxic content stated, *"That guy was blackmailing [me]. Because he [was claiming] I [was here] illegally. And [I cannot] report to police officers [although I wanted to]."*

Ignoring Toxic Content. Participants also practice nonintervention, where they choose to ignore toxic content (`FG1-9`, 10% of surveyed refugees). We note that given participants' confidence in social media use, ignoring toxic content did not stem from limited digital literacy. Instead, their decision not to intervene is influenced by one of three reasons. First, participants are concerned that responding could escalate a confrontation, causing more complications. Thomas et al. highlight how content creators, users who make digital content such as YouTube videos for income, avoid escalation as engaging can fuel further attacks [82]. Complementing this

study, and marginalization risk factors (negative treatment at societal level) outlined by Noel et al. [90], we find that participants attribute desire to avoid escalation to power dynamics between themselves and the attacker, who is likely to have citizenship / non-refugee status. To illustrate, one participant stated, *"[they] ignored [toxic content] because [they] did not want to make problems with residents of the country where [they] live."* Interestingly, we find this reason echoed among participants with intersectional identities. Similar to prior work showcasing how intersectionality shapes privacy responses [5], attacks against multiple identities are considered more severe and thus followed by nonintervention.

Second, nonintervention is motivated by the benefit of social media, with one participant sharing that social media is *"like medicine, sometimes it can be good, other times it can be bad."* These overwhelmingly positive aspects incentivize participants to *"turn a blind eye"* towards toxic content, seeing it as an unavoidable consequence of their identity/online presence and therefore believe it should be ignored.

Third, nonintervention is attributed to indifference (becoming accustomed to toxic content). Similar to sex workers [57], constant toxic content exposure can make participants apathetic towards attacks. One participant said that they chose not to block an attacker as *"[they] do not care about what they say"*, with another noting *"I ignore it. I don't want to waste my time with racist people."* Participants' decision to ignore is dependent on the severity of toxic content. For instance, one participant noted *"if the post is really outrageous, [I will respond]. Other content does not deserve a response"*.

Liaison participants admit that toxic content responses are largely absent from S&P guidelines/advice given to refugees. Current S&P advice for refugees outlines more traditional aspects of security (e.g., strong passwords, account sharing) [73]. Similarly, guidelines for social media focus on responsible use, e.g., one liaison participant working with refugees in a camp noted advice primarily involved avoiding pornography or minimizing online gaming. We posit that due to the absence of guidelines, participants do not intervene, feeling toxic content is unavoidable and requires no response.

Discouraging Toxic Content via Response. Participants' immediate responses are driven by their goal to discourage toxic content. Participants express how they are motivated to inform the platform and *"show their dissatisfaction with online hate"* (`FG1-9`). They intend to discourage toxic content by standing up to attackers (direct engagement, selective blocking) or instigating platform involvement (reporting).

Discouragement also indicates intent to protect community members. Similar to how toxic content against content creators can negatively impact their audiences [82], participants express how attacks may extend to others, e.g., *"[attackers] will not only go after me but [also my] close friends [and] family members"*. Just as victims of online hate and harassment attempt to stop attacks via social pleas [81], participants express dissatisfaction with online hate to prevent toxic content from affecting them and other refugees. Participants note a responsibility to prevent further attacks from affecting community members, e.g., *"I consider [my] friends and family members to, like, protect them from like, retribution"*.

6.3.2 Behavioral Changes Instigated By Toxic Content

Participants also enact behavioral changes in response to toxic content: (*a*) rigorous privacy measures, (*b*) activity in private groups, and (*c*) withdrawal from platforms.

Rigorous Privacy Measures. Noel et al. [90] note how at-risk communities often reconsider privacy measures, attempting to minimize digital footprints. We complement these findings as we discover that toxic content pushes participants to strengthen their privacy settings. Our participants state that restricting or redacting their personal identifiable information (PII) is the most common way they rethink their digital privacy presence (`FG1-9`), as PII can be used to track them. For instance, participants are wary of sharing visual content that can be used to identify them, e.g., their face picture or a photo of them at work. Liaisons also support this fact, saying that refugees' toxic online experiences result in the caution of sharing identifiable data. For instance, one liaison notes an incident where an attacker *"threatened a refugee after [they] posted [their] picture working at a local business."*

Refugee participants consider location data highly sensitive, a sentiment echoed by other populations (e.g., human trafficking survivors, undocumented immigrants [19, 42]). We find that participants' sensitivity towards location privacy is compounded as even generic location (e.g., city/district/state name) leaves them vulnerable. Given a region name, an attacker can associate the participant with a refugee camp or neighborhood with a high volume of refugees and subsequently use it to infer their fine-grained location.

However, we find that participants take extra precautions to obfuscate location data. Participants' caution extends beyond sharing explicit location data (e.g., street/building name) - liaison participants note that refugees limit access to information from which location can be inferred. We posit this attempts to strike a balance between privacy and utility - participants seeking to use social media still while preserving/obscuring their privacy-sensitive data. For instance, one liaison participant notes *"refugees do not post pictures ... near landmarks because it might be used against them."* Participants also note using privacy settings to delineate access to sensitive data on an individual basis. For instance, one participant noted *"If [it's for my] friends, I can share if strangers, so I can't, for safety reasons, and Yeah, that's about privacy"*.

Activity in Private Groups. We previously outlined (Section 6.1), how participants favor private groups that filter members. However, we find that *exclusive* and *increased* private group participation is a result of toxic content. This shift follows after a participant responds to a post, only to be tar-

geted by attacks. Private groups evolve from support sources (e.g., NGO coordination) to become participants' only form of online interaction. Liaison participants note inability to join some closed groups. Contrarily, closed groups with liaison participation emerge as a means to seek support and advice for refugees who are new to experiencing toxic content.

Withdrawal from Platforms. Participants also consider withdrawal from platform participation (e.g., deleting their account, or choosing a different but similar platform) when exposed to toxic content. For instance, one participant who experienced an attack *"deleted [their] old Facebook account [and] created a new one"* after *"hateful content was posted."*

Platform withdrawal as a response to online hate is common among demographics such as women [14] and even in specialized communities such as content creators [82]. There is an overlap with such communities in the impact of withdrawal (e.g., it can affect financial gain and platform enjoyment). However, refugee participants who withdraw are severed from integral support groups and the ability to reunify. Because of this, withdrawal is uncommon - only 10% of surveyed refugees practiced this behavioral change. Participants acknowledge preferring other behavioral changes, e.g., privacy measures, and moving towards private groups, as these changes allow them to still benefit from the platform.

Anonymity and Minimizing Vulnerability. Participants' behavioral changes focus on protecting themselves from future attacks via anonymity and minimizing vulnerability. Participants protect their anonymity by limiting other users' access to their PII via rigorous privacy settings, to protect their identity on social media. Anonymity may not reduce toxic content exposure – participants may still come across toxic content against refugees under an anonymous profile. However, it serves as a first line of defense to prevent escalation of attacks, e.g., repeated attacks or doxing. We also find participants' goal for anonymity is not solely focused on personal anonymity as they believe that protecting their own anonymity, in turn, protects the anonymity of their loved ones. For instance, one interviewed participant stated *"the whole point is to ... remain anonymous [otherwise] it will make it difficult for everyone [including] family from home country."*

Increased private group activity and withdrawal are motivated by an intent to minimize vulnerability and reduce toxic content exposure. For example, one participant who listed vulnerability-minimizing actions as responses to toxic content stated *"they don't like to get friendly on social media"* and *"try to avoid it"*. Here, participants note that limiting engagement on social media limits attack exposure. Interestingly, we note that withdrawal from the platform is the most severe method adopted by participants to minimize vulnerability. Participants share that withdrawal is the only foolproof method to eradicate interaction with toxic content attackers.

6.4 Barriers to Protective Practices

Although participants' protective actions are motivated by goals; they face barriers. Expectedly, digital literacy was not a barrier for our participants, given their self-reported confidence in using social media (Section 6.1). However, participants express two barriers related to trade-offs of protecting themselves from toxic content.

Privacy-Utility Trade-off. Similar to at-risk users, participants note basic needs as a barrier [90] to digital safety, with their unique necessities of social media becoming competing priorities that prevent S&P actions. Participants feel that strengthening their online privacy settings or moving to closed groups is a barrier depending on what these priorities are. For example, participants who depend on self-employment often rely on social media as a marketing tool for their services/products (e.g., *"they need to use social media to publicize their businesses"*). Additionally, as Geeng et al. [39] note, barriers may be impacted by intersecting identities (e.g., LGBTQ+). For instance, LGBTQ+ refugees who partake in advocacy groups, such as the Free Block 13 Campaign in Kakuma, rely on public exposure to gather support (e.g., *"Most people, I know, don't necessarily put Kakuma on their profile. And without that, you wouldn't necessarily be able to tell [preventing others from reaching out]"*. However, privacy and utility trade-offs inconvenience participants, who are forced towards more private and closed settings and forgo the use of social media. Similarly, withdrawal is difficult for participants who gain support via private groups.

Anonymity as a Barrier to Reunification. Although some restrict PII as a response to toxic content, some participants redact this PII to the point of anonymity (e.g., avoid using real names as usernames on social media platforms). However, this is a barrier for participants who need to remain identifiable to friends/family who seek to connect with them. Anonymity is especially challenging for participants who rely on social media for reunification - the rejoining of families separated by the refugee crisis. For instance, one participant expressed that their actions *"...the whole point is to like, remain anonymous ... [but] obviously makes it difficult for everyone. Family from [my] home [will not even know I] am fine."*

Finally, participants with no barriers attribute this to their lax implementation of toxic content protection mechanisms, as they use social media minimally.

6.5 Perspectives on Platform Involvement

Although prior work has argued for limiting digital-safety options [79], participants advocate for varying mitigation options, as more options provide tailored needs to individuals.

Labels can Empower. Warning labels can help in preventing attackers from posting toxic content while also protecting targets, as echoed in prior work [81]. Extending this, we find that labels can empower participants to contact platforms to either

express dissatisfaction or report an account. We also observe focus group participants actively advocate for warning labels because it can empower those who typically avoid reporting due to power dynamics (e.g., due to attacker's status). For instance, one participant who normally avoids reporting notes platforms should *"[tell us] this is ... toxic content [so] we can contact customer service"*, highlighting that labels are seen as "encouragement" to contact the platform.

Interestingly, one participant also expressed interest in labels for profiles (instead of posts), saying *"There could be some rating for every person ... if people [post] toxic content, you give them a low rating. So that when you visit [someone's] profile ... they have a percentage [score]."*, with other focus group participants agreeing that such an option would be beneficial. Here, we extend existing findings [81] by suggesting that presentations of labels can be updated to leverage existing information from user profiles (e.g., history of toxicity).

Censorship Shortcomings. Participants indicated support for stronger censorship of toxic posts, complementing other online communities' perceptions of toxic content moderation [82]. However, participants also outline shortcomings in existing censorship on platforms. To illustrate, one participant who experienced toxic content via online gaming stated they wanted better filtration of profane or offensive language. This participant expressed, *"[gaming service] should remove bad word(s) from the chat. Sometimes, they['ve] removed [it] already, but [miss some] words [because] it's in a different language."* "Different languages" not only references non-English languages but also "coded" hate words (e.g., slurs known amongst gaming communities or that target refugees). Participants attribute perceived shortcomings of existing moderation due to these nuances (a moderator may not necessarily recognize a slur). Although participants perceive censorship as flawed, it does help participants feel safer.

Distrust of Automation. Similar to suggestions for human moderation due to toxic content's contextual nature [90], participants' unique experiences result in a distrust of automated systems involved in reporting mechanisms. For example, one participant reported a post they felt was toxic against refugees but *"[were] quite surprised because [the system] said this is sensible content."* Although participants acknowledged a lack of familiarity with how automated systems worked, they shared that they would feel better if a human reviewed a report/complaint. For instance, one refugee who expressed an interest in technology and AI stated *"there could be more human checking ... on this toxic content ... relying on AI is not very applicable because it's not smart enough [to] detect these contents."* Though participants acknowledge that human involvement does not guarantee toxic content removal/censorship, participants place higher levels of trust in human moderation.

7 Discussion and Limitations

We now synthesize takeaways for combating toxic content against refugees and acknowledge our study's limitations.

7.1 Toxic Content Threat Models

It is important to consider at-risk communities' nuances in understanding threat models [57, 73, 75]. Participants' social media use cases, such as establishing private groups for support or public groups for advocacy (e.g., Free Block 13 Campaign) and dependency on reunification produce an intricate threat model. For instance, although strangers perpetrate attacks, personal attackers also exist, especially due to the ambiguity behind one's status. Participants also rely on social media platforms to network. Self-employed participants network for business, while others depend on members of local communities (e.g., neighbors), contributing to their vulnerability as these parties can turn out to be attackers.

Prior work highlights the importance of designing solutions that do not prevent user benefit of platform amenities (e.g., income generation, joining similar-interest social groups) [39, 82]. We extend this by arguing that existing attempts at designing inclusive threat models are incomplete. For example, platform guidelines often prohibit attacks against identities associated with different groups, which span from "protected characteristics" (e.g., race, religion) [50] to groups historically a target of abuse such as caste or sexual orientation [84]. However, platform guidelines do not consider how refugee-targeted attacks are interpreted as free speech - a concern raised by interviewed participants. The right to free speech is present in many countries (e.g., in the US, it refers to "the right to speak, write, and share ideas and opinions without facing punishment from the government" [23]) and consequently, free speech on social media is a widely debated issue [37]. Attackers often argue that any censorship violates this right, prompting lax moderation [12]. Similarly, existing privacy/anonymity features may be infeasible for some refugees - privacy and anonymity limit social media use and impede family reunification. Without knowledge of barriers, threat models may appear to have easy mitigation strategies (e.g., blocking, moving to private settings) when successfully implementing these solutions may be challenging.

We propose that social media platforms should have an open communication channel with refugee populations to ensure guidelines appropriately acknowledge toxic content threat models. For instance, knowledge of how refugee-related toxicity is interpreted as free speech is pivotal to reframing platform guidelines to account for this (e.g., appropriately delineating between free speech and toxicity against vulnerable populations). We also encourage future work to examine in-population differences within refugee communities, e.g., LGBTQ+ refugees, so platforms can develop a better understanding of toxic content threat models.

7.2 Detection and Mitigation Strategies

Despite guidelines against hateful content on social media platforms [44, 46, 49] and existing protective mechanisms to combat online hate, we find these guidelines are not grounded in refugees' experiences. Thus, toxic content persists on these platforms which we attribute to one of the following two reasons: inability (or delay) in detecting toxic content and lack of enforcement of platform guidelines.

Shortcomings of Detection. Work has explored automated detection of toxic content [20, 26, 30, 32, 60] and detecting attackers instead of their posts [16, 17, 33]. Community efforts, such as Perspective API [64], have also worked towards identifying online toxic content. However, automated mechanisms are often unable to capture the toxicity directed at specific groups (e.g., hateful hashtags against refugees). This can be mainly attributed to statistical or deterministic models used in these mechanisms not learning terms related to marginalized populations, varying cultural contexts across regions, and lacking representative datasets [81].

Despite Twitter disallowing attacks affecting marginalized groups [46], toxic hashtags against refugees (e.g., #NoRefugees, #Rapefugees) and other communities [9, 50] still exist. "Coded" hate words can also evade detection, as expressed by participants who are involved in gaming. However, we suggest that overcoming detection efforts' limitations is feasible through community collaboration. Prior work has shown that crowd-sourcing training data labels are often outsourced to specific regions [62]. We propose that social media platforms incorporate specific populations in enriching training data - platforms should include features for individuals/NGOs acutely aware of refugee-specific toxic content so that they can flag hashtags or coded slurs. Through this, developers of classification tools can design frameworks (e.g., annotation guides) to consider such nuances. This prevents imbalanced data and corresponding failure at detection which translates to participant distrust of automated detection mechanisms.

Building Towards Accessible Mitigation. Our study extends findings on the impracticality of one-size-fits-all approaches for marginalized communities [39], showing how participants have varying opinions on what mitigation works best and that unified strategies are impractical.

Participants advocate for labels against profiles (in lieu of per post). Similarly, some suggested moving from binary indicators toward numeric indicators that reflect severity. However, we recommend future work explore mitigation efforts in quantitative detail (e.g., via large-scale surveys) to fine-tune what exactly is shown in these labels/warnings. For instance, although warning labels embedded with an account's longitudinal information (e.g., previous violations) may benefit refugees, prior work shows that fewer/simpler digital safety options may prevent users from being overwhelmed [79, 92]. Thus, a large-scale study is required to decide what information can be presented to users to strike a balance between informing and not overwhelming them. Similarly, large-scale surveys are required to determine users' authority over labels (e.g., whether the ability to turn labels on/off is beneficial or what the default setting for the label should be).

Accessible mitigation should include refining automated censorship of toxicity. However, given automated detection's limitations and the contextual nature of toxic content, some participants prefer human moderation. Although it is unrealistic to manually parse every post/comment, human involvement in detection can bridge the gap that results from automation. Detection systems should continuously be updated to include community-driven information, such as toxic word lists from HateBase API [45] and regularly updated lists of hateful hashtags as suggested by recent work [81].

7.3 S&P Advice for Toxic Content

Prior work has exposed how access to guidelines and resources for dealing with online hate is imperative to online users such as creators [82]. We echo this, but note that the design of guidelines for refugees faces barriers. First, refugee organizations are in the early stages of developing S&P guidelines concerning toxic content. Second, advice given is intentionally limited to avoid burdening refugees beyond critical S&P, such as safe passwords. Third, guidelines should account for varying levels of digital literacy, refugee barriers, and their unique online use cases.

To overcome barriers, we suggest that S&P guidelines for refugees should include preemptive protective practices, e.g., secure distribution of group links to prevent infiltration, and removing location data from posts. Guidelines can also enumerate alternative mechanisms (e.g., when to avoid intervention and report, how to push back while maintaining privacy) that avoid impeding refugees' social media use.

Synthesizing guidelines is only the first step - guideline dissemination is also imperative to ensure refugees are exposed to advice. For instance, NGOs should incorporate mitigation strategies and information on toxic content threat models in classes often provided to refugees (e.g., digital literacy classes, citizenship classes). Refugee dependency on private groups makes groups appropriate sites to disseminate guidelines too. We also see a role for platforms in informing refugees. Guidelines, e.g., adjusting privacy to deal with toxic content, can be incorporated into website walkthroughs shown to refugees the first time they access the platform.

However, we recommend further large-scale interventions such as community workshops [76] to design and update proposed S&P guidelines. Guidelines should be iteratively improved with feedback from refugees to ensure that advice is actionable and does not burden refugees.

7.4 Limitations

Our study might be vulnerable to usual qualitative study constraints (e.g., observer bias, participant self-censorship). Additionally, prior work has outlined that online focus groups face barriers (e.g., technical difficulties, participant discomfort in front of the camera) [3, 10]. However, we conduct our best efforts to minimize these limitations' effects. Given that liaisons depend on recollections (observer bias), we survey and conduct focus groups with refugees who can comment on direct experiences. We interview liaisons for insights that refugees may be unwilling to share (due to self-censorship). We ensured initial/follow-up questions in focus groups were designed to elicit general responses first, with refugees providing detail when comfortable. We also conducted check-ins at the beginning, middle, and end of a session to ensure no technical difficulties were encountered.

We recruited participants from diverse regions of origin and settlement, employed a mixed-methods approach, and ceased recruitment upon our data's thematic saturation to improve generalizability. Our participants' experiences are grounded in the following commonalities. First, participants (or refugees liaison participants serve) use social media and online platforms. Second, they are exposed to toxic content in online spaces. Finally, their actions, goals, and barriers are influenced by their unique experiences as refugees. Our findings provide rich insight into toxic content's implications on refugees' security and privacy. We examine how in-population differences result in nuances in toxic content exposure (Section 6.2.3), but our participants skew Asian and Middle Eastern refugees who live in the Middle East or the US. Refugee participants also skewed towards those comfortable with using social media (high digital literacy) - those with lower literacy may have different actions and barriers. Additionally, our scope of social media may not account for additional toxic content experiences (e.g., when a refugee reads toxic comments on an online news website). Future work will examine differences in refugee toxic content exposure/action (1) across different settings (e.g., news websites vs streaming platforms), (2) in different countries and (3) with varying technical expertise in a demographically diverse in-depth study.

8 Conclusions

Toxic content attacks target refugees on social media, inflicting harm such as heightened anxiety and intent to silence. Through interviews, focus groups, and online surveys with refugees and liaisons, we discover participants' experiences shape S&P implications and responses. Their immediate responses and behavioral changes are motivated by goals such as minimizing vulnerability. Although actions have barriers, participants also outline mitigation efforts that may improve their online experiences. Our findings synthesize lessons to better refugees' ability to manage toxic content threats.

Acknowledgments

We thank our anonymous reviewers and shepherd for providing us with valuable feedback that helped improve our paper. We would also like to thank the participants for their generous time and contribution to our research. This work is supported by startup funding from Purdue University.

References

[1] Konstantin Aal, Marios Mouratidis, Anne Weibert, and Volker Wulf. Challenges of ci initiatives in a political unstable situation-case study of a computer club in a refugee camp. In *International Conference on Supporting Group Work*, 2016.

[2] Konstantin Aal, George Yerousis, Kai Schubert, Dominik Hornung, Oliver Stickel, and Volker Wulf. Come_in@ palestine: adapting a german computer club concept to a palestinian refugee camp. In *ACM International Conference on Collaboration across Boundaries: Culture, Distance & Technology*, 2014.

[3] Gail Adams-Hutcheson and Robyn Longhurst. 'at least in person there would have been a cup of tea': interviewing via skype. *Area*, 2017.

[4] Afghanistan situation. https://reporting.unhcr.org/afghansituation, 2022. [Online; accessed 13-August-2023].

[5] Tanisha Afnan, Yixin Zou, Maryam Mustafa, Mustafa Naseem, and Florian Schaub. Aunties, strangers, and the {FBI}: Online privacy concerns and experiences of {Muslim-American} women. In *Symposium on Usable Privacy and Security (SOUPS)*, 2022.

[6] Amanda Alencar. Refugee integration and social media: A local and experiential perspective. *Information, Communication & Society*, 2018.

[7] Asam Almohamed and Dhaval Vyas. Designing for the marginalized: A step towards understanding the lives of refugees and asylum seekers. In *ACM Conference Companion Publication on Designing Interactive Systems*, 2016.

[8] Nazanin Andalibi, Oliver L Haimson, Munmun De Choudhury, and Andrea Forte. Understanding social media disclosures of sexual abuse through the lenses of support seeking and anonymity. In *CHI Conference on Human Factors in Computing Systems*, 2016.

[9] Antisemitic posts are rarely removed by social media companies. https://www.npr.org/2021/08/02/1023819435/antisemitic-posts-are-rarely-removed-by-social-media-companies-a-study-finds. [Online; accessed 30-August-2023].

[10] Mandy M Archibald, Rachel C Ambagtsheer, Mavourneen G Casey, and Michael Lawless. Using zoom videoconferencing for qualitative data collection: perceptions and experiences of researchers and participants. *International Journal of Qualitative Methods*, 2019.

[11] Carlos Arcila-Calderón, David Blanco-Herrero, Maximiliano Frías-Vázquez, and Francisco Seoane-Pérez. Refugees welcome? online hate speech and sentiments in twitter in spain during the reception of the boat aquarius. *Sustainability*, 2021.

[12] Twitter takeover. https://www.brookings.edu/articles/why-is-elon-musks-twitter-takeover-increasing-hate-speech/, 2022. [Online; accessed 1-August-2023].

[13] Jenna Burrell and Ken Anderson. I have great desires to look beyond my world': trajectories of information and communication technology use among ghanaians living abroad. *New Media & Society*, 2008.

[14] Kalyani Chadha, Linda Steiner, Jessica Vitak, and Zahra Ashktorab. Women's responses to online harassment. *International Journal of Communication*, 2020.

[15] Rahul Chatterjee, Periwinkle Doerfler, Hadas Orgad, Sam Havron, Jackeline Palmer, Diana Freed, Karen Levy, Nicola Dell, Damon McCoy, and Thomas Ristenpart. The spyware used in intimate partner violence. In *IEEE Symposium on Security and Privacy (SP)*, 2018.

[16] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. Hate is not binary: Studying abusive behavior of# gamergate on twitter. In *ACM Conference on Hypertext and Social Media*, 2017.

[17] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. Mean birds: Detecting aggression and bullying on twitter. In *ACM on Web Science Conference*, 2017.

[18] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. Measuring# gamergate: A tale of hate, sexism, and bullying. In *International Conference on World Wide Web Companion*, 2017.

[19] Christine Chen, Nicola Dell, and Franziska Roesner. Computer security and privacy in the interactions between victim service providers and human trafficking survivors. In *USENIX Security Symposium*, 2019.

[20] Hao Chen, Susan McKeever, and Sarah Jane Delany. The use of deep learning distributed representations in the identification of abusive text. In *International AAAI Conference on Web and Social Media*, 2019.

[21] Shira Chess and Adrienne Shaw. A conspiracy of fishes, or, how we learned to stop worrying about# gamergate and embrace hegemonic masculinity. *Journal of Broadcasting & Electronic Media*, 2015.

[22] Cloud translation api. https://cloud.google.com/translate, 2021. [Online; accessed 14-November-2022].

[23] freedom of speech. https://www.law.cornell.edu/wex/freedom_of_speech, 2021. [Online; accessed 15-August-2023].

[24] Kimberlé W Crenshaw. *On intersectionality: Essential writings*. The New Press, 2017.

[25] Negin Dahya and Sarah Dryden-Peterson. Tracing pathways to higher education for refugees: the role of virtual support networks and mobile phones for women in refugee camps. *Comparative Education*, 2017.

[26] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *International AAAI Conference on Web and Social Media*, 2017.

[27] Rianne Dekker, Godfried Engbersen, Jeanine Klaver, and Hanna Vonk. Smart refugees: How syrian asylum migrants use social media information in migration decision-making. *Social Media+ Society*, 2018.

[28] Tibor Dessewffy and Zsofia Nagy. Born in facebook: The refugee crisis and grassroots connective action in hungary. *International Journal of Communication*, 2016.

[29] Jayati Dev, Pablo Moriano, and L Jean Camp. Lessons learnt from comparing {WhatsApp} privacy concerns across saudi and indian populations. In *Symposium on Usable Privacy and Security (SOUPS)*, 2020.

[30] Karthik Dinakar, Roi Reichart, and Henry Lieberman. Modeling the detection of textual cyberbullying. In *International AAAI Conference on Web and Social Media*, 2011.

[31] S Dryden-Peterson, N Dahya, and D Douhaibi. How teachers use mobile phones as education tools in refugee camps. *Education and Information Technologies*, 2017.

[32] Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *International AAAI Conference on Web and Social Media*, 2018.

[33] Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. Peer to peer hate: Hate speech instigators and their targets. In *International AAAI Conference on Web and Social Media*, 2018.

[34] Julia R Fernandez and Jeremy Birnholtz. " i don't want them to not know" investigating decisions to disclose transgender identity on dating platforms. *ACM on Human-Computer Interaction (CSCW)*, 2019.

[35] Free block 13. https://www.facebook.com/groups/freeblock13/, 2021. [Online; accessed 13-August-2023].

[36] Diana Freed, Jackeline Palmer, Diana Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. "a stalker's paradise" how intimate partner abusers exploit technology. In *CHI Conference on Human Factors in Computing Systems*, 2018.

[37] freedom of speech. https://www.freedomforum.org/free-speech-on-social-media/, 2021. [Online; accessed 15-August-2023].

[38] Damjan Fujs, Anže Mihelič, and Simon LR Vrhovec. The power of interpretation: Qualitative methods in cybersecurity research. In *International Conference on Availability, Reliability and Security*, 2019.

[39] Christine Geeng, Mike Harris, Elissa Redmiles, and Franziska Roesner. Like lesbians walking the perimeter: Experiences of U.S. LGBTQ+ folks with online security, safety, and privacy advice. In *USENIX Security Symposium*, 2022.

[40] Marie Gillespie, Ampofo Lawrence, Margaret Cheesman, Becky Faith, Evgenia Illiou, Ali Issa, Souad Osseiran, and Dimitris Skleparis. Mapping refugee media journeys: Smartphones and social media networks. *ResearchReport*, 2016.

[41] Marie Gillespie, Souad Osseiran, and Margie Cheesman. Syrian refugees and the digital passage to europe: Smartphone infrastructures and affordances. *Social Media + Society*, 2018.

[42] Tamy Guberek, Allison McDonald, Sylvia Simioni, Abraham H Mhaidli, Kentaro Toyama, and Florian Schaub. Keeping a low profile? technology, risk and privacy among undocumented immigrants. In *CHI Conference on Human Factors in Computing Systems*, 2018.

[43] How social media's toxic content sends teens into 'a dangerous spiral. https://www.hsph.harvard.edu/news/features/how-social-medias-toxic-content-sends-teens-into-a-dangerous-spiral/, 2021. [Online; accessed 13-August-2023].

[44] Hate speech. https://transparency.fb.com/policies/community-standards/hate-speech/, 2023. [Online; accessed 13-August-2023].

[45] Hatebase. https://hatebase.org/about, 2021. [Online; accessed 13-August-2023].

[46] Hateful conduct policy. https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy, 2023. [Online; accessed 13-August-2023].

[47] Sam Havron, Diana Freed, Rahul Chatterjee, Damon McCoy, Nicola Dell, and Thomas Ristenpart. Clinical computer security for victims of intimate partner violence. In *USENIX Security Symposium*, 2019.

[48] Gabriel Emile Hine, Jeremiah Onaolapo, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Riginos Samaras, Gianluca Stringhini, and Jeremy Blackburn. Kek, cucks, and god emperor trump: A measurement study of 4chan's politically incorrect forum and its effects on the web. In *International AAAI Conference on Web and Social Media*, 2017.

[49] Instagram community guidelines. https://help.instagram.com/477434105621119, 2023. [Online; accessed 13-August-2023].

[50] 11,696 examples of how hate thrives on social media. https://www.nytimes.com/2018/10/29/technology/hate-on-social-media.html. [Online; accessed 30-August-2023].

[51] Intersectionality. https://en.wikipedia.org/wiki/Intersectionality, 2023. [Online; accessed 30-August-2023].

[52] Ramona Kreis. # refugeesnotwelcome: Anti-refugee discourse on twitter. *Discourse & Communication*, 2017.

[53] Linda Leung. Telecommunications across borders: Refugees' technology use during displacement. *Telecommunications Journal of Australia*, 2010.

[54] Carleen Maitland and Ying Xu. A social informatics analysis of refugee mobile phone use: A case study of za'atari syrian refugee camp. In *Conference on Communication, Information and Internet Policy (TPRC)*, 2015.

[55] Jay Marlowe. Refugee resettlement, social media and the social organization of difference. *Global Networks*, 2020.

[56] 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 *CHI Conference on Human Factors in Computing Systems*, 2017.

[57] Allison McDonald, Catherine Barwulor, Michelle L Mazurek, Florian Schaub, and Elissa M Redmiles. " it's stressful having all these phones": Investigating sex workers' safety goals, risks, and practices online. In *USENIX Security Symposium*, 2021.

[58] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *ACM on Human-Computer Interaction (CSCW)*, 2019.

[59] Impact of online hate. https://mediasmarts.ca/online-hate/impact-online-hate, 2022. [Online; accessed 13-August-2023].

[60] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In *International Conference on World Wide Web (WWW)*, 2016.

[61] Borke Obada-Obieh, Lucrezia Spagnolo, and Konstantin Beznosov. Towards understanding privacy and trust in online reporting of sexual assault. In *Symposium on Usable Privacy and Security (SOUPS)*, 2020.

[62] Openai outsourced data labeling. https://www.datanami.com/2023/01/20/openai-outsourced-data-labeling-to-kenyan-workers-earning-less-than-2-per-hour-time-report/. [Online; accessed 30-August-2023].

[63] Nazan Öztürk and Serkan Ayvaz. Sentiment analysis on twitter: A text mining approach to the syrian refugee crisis. *Telematics and Informatics*, 2018.

[64] Google jigsaw. https://www.perspectiveapi.com/#/, 2021. [Online; accessed 5-August-2023].

[65] Shruti Phadke and Tanushree Mitra. Many faced hate: A cross platform study of content framing and information sharing by online hate groups. In *CHI Conference on Human Factors in Computing Systems*, 2020.

[66] Refugee situation. https://www.unhcr.org/refugee-statistics/, 2023. [Online; accessed 30-August-2023].

[67] Jake Reichel, Fleming Peck, Mikako Inaba, Bisrat Moges, Brahmnoor Singh Chawla, and Marshini Chetty. 'i have too much respect for my elders' understanding south african mobile users' perceptions of privacy and current behaviors on facebook and whatsapp. In *USENIX Security Symposium*, 2020.

[68] Rebecca S Robinson. Purposive sampling. *Encyclopedia of Quality of Life and Well-being Research*, 2014.

[69] Jon Roozenbeek and Sander Van Der Linden. The fake news game: actively inoculating against the risk of misinformation. *Journal of Risk Research*, 2019.

[70] 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 *CHI Conference on Human Factors in Computing Systems*, 2019.

[71] 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*, 2018.

[72] Morgan Klaus Scheuerman, Stacy M Branham, and Foad Hamidi. Safe spaces and safe places: Unpacking technology-mediated experiences of safety and harm with transgender people. *ACM on Human-computer Interaction (CSCW)*, 2018.

[73] 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.

[74] 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 *USENIX Security Symposium*, 2022.

[75] 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 *USENIX Security Symposium*, 2022.

[76] Julia Slupska, Scarlet Dawson Dawson Duckworth, Linda Ma, and Gina Neff. Participatory threat modelling: Exploring paths to reconfigure cybersecurity. In *Extended Abstracts of CHI Conference on Human Factors in Computing Systems*, 2021.

[77] How social media's toxic content sends teens into 'a dangerous spiral'. https://www.hsph.harvard.edu/news/features/how-social-medias-toxic-content-sends-teens-into-a-dangerous-spiral/, 2022. [Online; accessed 30-August-2023].

[78] What doctors wish patients knew about social media's toxic impact. https://www.ama-assn.org/delivering-care/population-care/what-doctors-wish-patients-knew-about-social-media-s-toxic-impact, 2022. [Online; accessed 30-August-2023].

[79] Brian Stanton, Mary F Theofanos, Sandra Spickard Prettyman, and Susanne Furman. Security fatigue. *It Professional*, 2016.

[80] Reem Talhouk, Syed Ishtiaque Ahmed, Volker Wulf, Clara Crivellaro, Vasilis Vlachokyriakos, and Patrick Olivier. Refugees and hci sig: The role of hci in responding to the refugee crisis. In *Extended Abstracts of CHI Conference on Human Factors in Computing Systems*, 2016.

[81] Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley, Deepak Kumar, et al. Sok: Hate, harassment, and the changing landscape of online abuse. In *IEEE Symposium on Security and Privacy (SP)*, 2021.

[82] 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 *CHI Conference on Human Factors in Computing Systems*, 2022.

[83] Serdar Tuncer et al. Online hate on youtube: Anti-immigrant rhetoric against syrian refugees in canada and turkey. *Humanities Commons*, 2020.

[84] Hateful conduct. https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy, 2021. [Online; accessed 15-August-2023].

[85] Ukrainian refugees. https://www.bbc.com/news/world-60555472, 2022. [Online; accessed 30-August-2023].

[86] What is a refugee. https://www.unhcr.org/what-is-a-refugee, 2023. [Online; accessed 13-August-2023].

[87] Maximiliano Frías Vázquez and Francisco Seoane Pérez. Hate speech in spain against aquarius refugees 2018 in twitter. In *International Conference on Technological Ecosystems for Enhancing Multiculturality*, 2019.

[88] Jessica Vitak, Kalyani Chadha, Linda Steiner, and Zahra Ashktorab. Identifying women's experiences with and strategies for mitigating negative effects of online harassment. In *ACM Conference on Computer Supported Cooperative Work and Social Computing*, 2017.

[89] Ashley Marie Walker, Yaxing Yao, Christine Geeng, Roberto Hoyle, and Pamela Wisniewski. Moving beyond 'one size fits all': Research considerations for working with vulnerable populations. *Interactions*, 2019.

[90] 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.

[91] Mark Warner, Andreas Gutmann, M Angela Sasse, and Ann Blandford. Privacy unraveling around explicit hiv status disclosure fields in the online geosocial hookup app grindr. *ACM on Human-Computer Interaction (CSCW)*, 2018.

[92] 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 *CHI Conference on Human Factors in Computing Systems*, 2023.

[93] Ying Xu and Carleen Maitland. Communication behaviors when displaced: a case study of za'atari syrian refugee camp. In *International Conference on Information and Communication Technologies and Development*, 2016.

[94] George Yerousis, Konstantin Aal, Thomas von Rekowski, David W Randall, Markus Rohde, and Volker Wulf. Computer-enabled project spaces: Connecting with palestinian refugees across camp boundaries. In *CHI Conference on Human Factors in Computing Systems*, 2015.

[95] Savvas Zannettou, Joel Finkelstein, Barry Bradlyn, and Jeremy Blackburn. A quantitative approach to understanding online antisemitism. In *International AAAI Conference on Web and Social Media*, 2020.

[96] Adam G Zimmerman and Gabriel J Ybarra. Online aggression: The influences of anonymity and social modeling. *Psychology of Popular Media Culture*, 2016.

A Toxic Content Categories

We provide brief descriptions for toxic content subcategories in Table 1.

Table 1: Overview of Toxic Content Categories.

Toxic Content Type	Description
bullying	seek to harm or intimidate or coerce an individual perceived as vulnerable
trolling	intentionally provoke someone/group of people with inflammatory remarks
hate speech	contain abusive or threatening content that expresses prejudice targeting a group of people based on their race, gender, political/ideological affiliation, religion or a similar property
profane or offensive language	contain profane or offensive language (e.g., showing lack of respect to someone's religious beliefs, cursing, swearing, expletives, culturally offensive content)
threats of violence	content that physically threatens someone
purposeful embarrassment	content that tries to purposely embarrass someone
incitement	provokes unlawful behavior or urge someone to behave unlawfully
sexual harassment	sexually harasseses someone (e.g., Unwelcome sexual advances, requests for sexual favors)
unwanted explicit content	contains unwanted explicit content (e.g., sexting, violent and adult content)

B User Study Details

We provide our initial questions for both liaison semi-structured interviews and focus groups with refugees.

B.1 Interview Questions for Liaisons

We detail questions used in our interviews with refugee liaisons (Section 5.3). We note that our interviews were semi-structured, and the listed questions were only starting points of conversation, with follow-up questions asked based on participant responses.

1. What is your job title and role?
2. Can you tell us about the refugees you work with? You can provide any information you are comfortable sharing such as nationality, age range or country of origin. (We ask that you not use names or identifiable information.)
3. Which online platforms/apps do they use?
4. Have you ever witnessed/heard any toxic content directed at refugees you have worked with online?
5. Can you describe what you have witnessed/heard in your experiences of this kind?
6. When did this occur and on which platforms?
7. Was the behavior you described a result of any online campaign against refugees?
8. Can you list some reasons for posting this toxic content?
9. What are the impacts of online toxic content on social media usage of refugees?
10. What are the impacts of online toxic content on refugees' daily life?
11. Can you tell us about safety precautions taken by refugees in order to avoid this toxic content if there are any?
12. Do you give any advice to your clients in terms of social media usage? Do you have any documents/instructions containing this advice?
13. Assuming you work with refugees from multiple countries, are your observations any different for different countries?
14. Are you aware of any ongoing campaigns/attempts to discourage or stop posting hateful content?

B.2 Focus Group with Refugees

We detail questions used in our focus groups with refugees (Section 5.3). Similar to our semi-structured interviews with liaisons, these questions were only starting points, with follow-up questions based on refugee responses and discussion between refugees within focus groups. Questions were purposefully general and broad so as not to force refugees to detail information or experiences they were not comfortable sharing.

1. Can you please tell us about yourselves?
2. How often do you use social media? Which platforms do you use?
3. What is your main purpose of social media use?
4. What does online/digital security and privacy mean to you?
5. Has any toxic content/ online hate ever happened to you, personally online?

6. Can you describe what you have witnessed in your experiences of this kind?
7. In which of the following online environments did your experience(s) occur?
8. Did you use translation services to understand this content? If so, please specify what you used.
9. Do you think any of these experiences are a result of your race or ethnicity, gender, sexual orientation, political views, religion?
10. Thinking of the person or people involved, how did you know them?
11. Did you respond to this behavior or ignore it and why did you prefer to do so?
12. After the experience(s), did you continue to use the platform?
13. After the experience(s), did you lean towards other applications/platforms?
14. When posting online content, do you share your location or pictures? If not, why?
15. What are your (security/privacy) goals when adopting these practices?
16. Are any of the security practices you rely on barriers for you?
17. Is there anything else you would like to share about the precautions you take in order to avoid online toxic content?
18. Regarding toxic content, do you have any suggestions on how social media companies can improve the social media experience?