Article

SM+S
social media + society

Social Media + Society
January-March 2023: 1–12

DOI: 10.1177/20563051231157297
journals.sagepub.com/home/sms
SAGE

Online Harassment: Assessing Harms and Remedies

Sarita Schoenebeck, Cliff Lampe, and Penny Triệu

Abstract

Online harassment refers to a wide range of harmful behaviors, including hate speech, insults, doxxing, and non-consensual image sharing. Social media platforms have developed complex processes to try to detect and manage content that may violate community guidelines; however, less work has examined the types of harms associated with online harassment or preferred remedies to that harassment. We conducted three online surveys with US adult Internet users measuring perceived harms and preferred remedies associated with online harassment. Study 1 found greater perceived harm associated with non-consensual photo sharing, doxxing, and reputational damage compared to other types of harassment. Study 2 found greater perceived harm with repeated harassment compared to one-time harassment, but no difference between individual and group harassment. Study 3 found variance in remedy preferences by harassment type; for example, banning users is rated highly in general, but is rated lower for non-consensual photo sharing and doxxing compared to harassing family and friends and damaging reputation. Our findings highlight that remedies should be responsive to harassment type and potential for harm. Remedies are also not necessarily correlated with harassment severity—expanding remedies may allow for more contextually appropriate and effective responses to harassment.

Keywords

online harassment, online harms, online remedies, content moderation

Introduction

Online harassment refers to a range of online behaviors from insults to hate speech to threatening physical violence (Duggan, 2017; Pater et al., 2016). Social media platforms rely on a set of predetermined rules for determining what kinds of content violates community guidelines or not, which are then used to govern millions or billions of posts per day across platforms. These rules are enacted by a process called content moderation that relies on human workers who manually label content in combination with computational techniques that detect and enforce rules algorithmically (Gillespie, 2018; Roberts, 2019). If content is found to violate guidelines, the violating content may be removed, and the person who posted it may be warned or banned (Gillespie, 2018; Pater et al., 2016; Roberts, 2019).

Despite advances in content moderation, social media platforms have struggled to effectively moderate online behavior (Goldberg, 2019; Stevenson, 2018; York, 2021). Users engage in hate speech, insults, threats, embarrassment, non-consensual sharing of images, and doxxing, and many of these behaviors elude moderation system filters. Some problems are challenges with context—decisions about individual pieces of content applied in one context may fail if applied to different contexts, languages, cultures, and so on (Douek, 2020; York, 2021). Other problems are with scale—content moderation relies on human workers who are typically paid low wages to review large amounts of content quickly, some of which is traumatizing to view (e.g., child abuse photos) (Roberts, 2019). Removing content also raises concerns about censorship—social media platforms have tremendous power to determine what billions of users can or cannot say on their platforms with little oversight or accountability for that power (Citron, 2017; Kaye, 2019; York, 2021). Finally, harassing behaviors are nuanced and contextualized; they may be private via direct message or public on news feeds or streams. They may also be one-on-one, in groups, or networked—where coordinated attacks are orchestrated against an individual or community (Gray, 2020; Marwick, 2021).

University of Michigan, USA

Corresponding Author:
Sarita Schoenebeck, University of Michigan, Ann Arbor, MI 48109, USA.
Email: yardi@umich.edu

While a substantial body of work has examined the work of paid content moderators, volunteer moderators, and of governance policies, there has been less work focused on mapping perceived harms associated with online harassment and remedies for those harms. The current article expands recent work assessing the severity of harmful content online but which does not focus specifically on online harassment (Jiang et al., 2021; Scheuerman et al., 2021). It also builds on work focused on expanding remedies but which does not disambiguate preferences based on types of online harassment (Bridy, 2018; Goldman, 2021; Schoenebeck et al., 2020).

We conducted three online studies to measure perceptions of harm and preferences for remedies. The first study examined perceived sexual, physical, and psychological harm associated with 11 diverse types of online harassment. Drawing on results from Study 1, we selected four types of online harassment for use in Studies 2 and 3. Study 2 examined the perceived harms associated with repeated versus one-time harassment and with individual versus group harassment for four harassment types. Study 3 examined preferred remedies associated with four harassment types, including both existing responses and potential new ones derived from prior work and theory. We discuss contextually sensitive remedies for online harassment, which consider the nature of the harassment and the potential for harm. We advise that an expanded suite of possible remedies can shift platforms' dependence on banning and removal when other responses may be more effective or appropriate approaches to governance.

Related Work

Types of Online Harassment

Online harassment is experienced in most regions of the world: just under half of Internet users have experienced online abuse in some form across 22 countries globally, although prevalence by country ranges from about 20% in Japan to 72% in Kenya ("SoK: Hate, Harassment, and the Changing Landscape of Online Abuse," 2021). In the United States, about 40%–50% of adult internet users report experiencing harassment (Duggan, 2017; Lenhart et al., 2016). Online harassment is a broad term that covers many online experiences and there is no commonly agreed on term for the range of behaviors it might encompass. Duggan (2017) defined six types of harassment in their report: offensive name-calling, purposeful embarrassment, physical threats, sustained harassment, stalking, and sexual harassment. Lenhart et al. (2016) defined three major types of harassment in theirs: direct harassment, experienced by 36% of their sample, was defined as "things that people do directly to one another" and included offensive name-calling, physical threats, purposeful embarrassment, sexual harassment, sustained harassment, identify theft, spreading rumors and stalking, among a few others. In psychology and counseling literature, the term "cyberbullying" has been commonly used, though it is often discussed in the context of harassment that occurs between young people (Kowalski et al., 2014; Slonje & Smith, 2008). In that literature, there has been a long-standing debate about how to define the term "cyberbullying" and definitions often include behaviors like victimization, offending, and nontraditional peer aggression (Hinduja & Patchin, 2010; Langos, 2015).

Women are more likely to be victims of gender-based harassment, though men can be victims as well (Aziz, 2015; Citron, 2022; Goldberg, 2019; UNESCO, 2021). People who are non-binary or other gender minorities are likely to be victims as well, though less studied currently. Gender-based harassment can include intimate partner violence (IPV), non-consensual image sharing, digital dating abuse, and misogynoir, among many other types (Bailey, 2021; Dragiewicz et al., 2018; Reed et al., 2017; Rocha-Silva et al., 2021; Tseng et al., 2020). Although these types of harassment differ in their characteristics, they share a desire to control, surveil, and monitor a person. Gendered harassment is difficult to address technically because it often occurs in private interpersonal interactions (e.g., direct messages) and can be easily reproduced across sites (e.g., websites specifically designed for sharing "revenge porn"). It is also hard to address socially or legally because of deep-seated global societal inequities that may not to see gender-based violence as a first-order concern (Heise, 1994; World Health Organization, 2021).

Other definitions of online harassment consider power differentials and persistence (Kazerooni et al., 2018). Cheng et al. (2017) use the broad behavior of trolling to define what they call "antisocial behaviors" and operationalize trolling by the emotional valence of the response (Cheng et al., 2017). Cheng et al. (2015) similarly define antisocial behavior based on the outcomes of a message, rather than the message itself, including writing quality over time, community responses, and similar features (Cheng et al., 2015). Blackwell et al. (2018) describe harassment as a broad set of behaviors, which includes flaming (i.e., posting insults), doxxing (i.e., posting personal information like home address with the intent of inducing harm), impersonation (i.e., creating a fake profile pretending to be someone else) and public shaming (Blackwell et al., 2018). Across the definitions and types of online harassment, harmful behaviors are characterized by aggression, intentionality, and harm (Blackwell et al., 2017; Duggan, 2017; Langos, 2015; Lenhart et al., 2016).

For the current study, we created a list of 11 harassment types using an iterative and qualitative approach based on prior literature. To do this, we started with the harassment types used by Duggan, 2017 and Lenhart et al. (2016), including physical threats, stalking, sustained harassment, sexual harassment, offensive name-calling, and purposeful embarrassment (Duggan, 2017; Lenhart et al., 2016). We then listed harassment types from other scholarship cited above and used whiteboards, notepads, and shared digital

documents to list a wide range of harassment types then discussed them among the coauthors. We used two predominant criteria to reduce the list of harassment types down for Study 1. The first criterion we used was prevalence in prior work—we focused on types of harassment, which were widely reported and observed in prior literature and case studies. The second criterion we used was diversity of harassment experiences—we wanted to capture a range of experiences that might yield different perceived harms and preferred remedies. We also discussed the types with colleagues who have expertise in online harassment and social media for further feedback and refinement. Our process of developing a set of types of harm resulted in 11 final types for Study 1.

Harm and Online Harassment

Harm has diverse meanings that vary with use and context. In legal contexts, harm refers to loss or damage to a person's right, property, or physical or mental well-being. Some harms are small and reparable, such as theft of a bicycle; others, such as loss of reputation or health, are irreparable and cannot be adequately compensated. Harms associated with technology have ranged from data harms to privacy harms to gendered harms (Citron, 2022; Scheuerman et al., 2021; Schoenebeck & Blackwell, 2021; Solove & Citron, 2017).

Social media platforms can enable intentional harms (e.g., doxxing a journalist because she wrote something a perpetrator did not like) or unintentional (e.g., using language on Twitter that may be exclusive of disabled people). Harm is also distinct from the act that causes the harm. An act may be perceived as minor for the perpetrator but particularly harmful to its recipient (e.g., reading a tweet that triggers retraumatization)—and vice versa. In addition, a single harmful experience can be differentially traumatic to different types of people. Different experiences can be differentially traumatic to different types of people (Bloom, 1999; Harms, 2015), but common characteristics include a sense of loss of control, negative emotional reactions, and critical assessments of self. These experiences can subsequently cause mood disorders, anxiety disorders, eating disorders, social phobias, trouble sleeping, dysthymia and a range of other negative psychological and physical experiences (Bloom, 1999; Harms, 2015).

Jiang et al. (2021) developed a taxonomy of online topics that cause harm that range from self-harm to regulated goods. Scheuerman et al. (2021) propose four types of harm in their work categorizing harmful online behaviors broadly—physical, emotional, relational, and financial. Citron (2022) has proposed a right to intimate privacy to protect against surveillance and misuse of private intimate and sexual data, including sexual content that is used to harass victims online. Aligned with our focus on interpersonal harms, we drew from the three types of harm prominently used by the World Health Organization (2021), Centers for Disease Control (2021), and myriad clinical studies and human rights work: physical, sexual, and psychological harms. Physical harm involves bodily injury or changes in environments that can relate to bodily safety such as lack of access to housing. Sexual harm relates to sexual abuse, including rape. Psychological harm involves mental and emotional states, and is likely co-occur with the other two categories. Two of these types overlap with Scheuerman et al.—physical harm and emotional/psychological harm—while sexual harm introduces a different dimension documented by Citron and others. There are other types of harm, including financial, economic, and reproductive harms that we did not focus on. Our first research question is as follows:

RQ1. What are the perceived harms associated with different types of online harassment?

While perceived harms may vary by harassment type, they can also vary based on the nature of the harassment. Online harassment may be especially severe in harms if it is repeated—many victims of online harassment, especially women, have documented the ongoing, repeated harassment they have experienced as part of being online (Goldberg, 2019; Gray, 2012; Massanari, 2017; UNESCO, 2021; Usher et al., 2018). Harassment can also be perpetuated by groups of people, either by relatively uncoordinated crowds where people act independently but en masse or through deliberate brigading where multiple people consciously coordinate to harass a target (Blackwell et al., 2017; Kazerooni et al., 2018; Marwick, 2021; Marwick et al., 2019; Stroud, 2016). Given the severity of prior experiences of harassment that are repeated or perpetuated by a group, we hypothesize that these may be more harmful than one-time or individual harassment. Here, we propose two hypotheses:

H1a. Perceived harm is greater when online harassment is repeated compared to when it is one-time.

H1b. Perceived harm is greater when online harassment is perpetuated by a group rather than an individual.

Harassment is experienced differently by different types of people. As Duggan (2017) shows, people 18–29 in the United States are much more likely than older people to experience online harassment, which could be correlated with a higher likelihood to participate in online activities. Lenhart et al. (2016) find multiple differences in how demographics changed the frequency of harassment, in addition to the effects of age described. They find that Black people were more likely to receive multiple types of harassment. Lenhart et al. (2016) found that over half of victims of harassment reported being worried, scared, angry and/or annoyed by their experience, and that these experiences made them hesitant to engage further. They found 27% of people overall reported that they do not post because of this fear, but 41% of women aged 15–29 reported this effect.

Scholar Moya Bailey (2021) created the term "misogynoir" to describe anti-Black misogyny that is often perpetuated in online spaces (Gray, 2012, 2020) In gaming, women of color experience griefing (i.e., persistent harassment to decrease game enjoyment), flaming, and hate speech (Gray, 2012). Online racism includes hate crimes and discrimination, and can also include behaviors such as micro-insults, invalidation of experiences of people of color, sharing texts and videos describing violence against people of color, and appropriation as types of online aggression (Stewart et al., 2019). Lenhart et al. (2016) also find that lesbian, gay, and bisexual (LGB) people are more likely to experience multiple forms of harassment, including offensive name-calling, sexual harassment, stalking, and more. LGB respondents reported being twice as likely (57% vs 26%) as heterosexual respondents to not post due to a fear of harassment. As mentioned above, numerous studies in the United States and globally show that women in many parts of the world experience severe online harassment, with associated stigma, family shame, violence, and other negative outcomes (Aziz, 2015; Lirri, 2015; Online Violence & Internet Harassment of Women, 2015; Pskowski, 2017; UNESCO, 2021). Studies also suggest that people with disabilities can experience online harassment and cyberbullying based on their identities (Alhaboby et al., 2016). Given these prior reports, we anticipate that perceived harms associated with online harassment may vary by identity. Thus, our second research question is as follows:

RQ2. How does perceived harm associated with types of online harassment vary by identity?

Online Harassment Remedies

Content removal and user sanctions or bans are the predominant mechanisms for sanctioning inappropriate content in many social media platforms; however, they offer a limited set of remedies for a wide range of social problems (Goldman, 2021). Removing content or banning users can be a quick fix to a harassment problem, but it may fail to remediate underlying behaviors—offenders have no opportunity for accountability while victims have no opportunity to experience justice or repair. Drawing on the concept of remedies from legal scholarship (Goldman, 2021), this project expands recent work by Schoenebeck et al. (2020) examining preferred responses to online harassment. The current study refines the set of remedies proposed in Schoenebeck et al. and explores whether those remedies vary by harassment type.

Some research suggests that removing harmful content and the users who post it can improve the overall quality of a platform, though efficacy is difficult to measure due to the technical challenges of tracking users across communities and platforms (Chandrasekharan et al., 2017). Despite substantial advances in content moderation, approaches to governance have remained somewhat narrowly focused on perpetrators of bad behavior and how to respond to that behavior (Bradford et al., 2019; Chandrasekharan et al., 2018; Goldman, 2021; Jhaver et al., 2019; Matias, 2019; Pater et al., 2016; Perez, n.d.).

When people are the targets of online harassment, they are often given little opportunity to be heard or to experience repair (Blackwell et al., 2017; Schoenebeck et al., 2020). They may receive a message that the content did or did not violate community guidelines, but they are given little ability to have a voice in their experience beyond that. In addition, banning users and removing content may effectively removing harmful behaviors and people but it is not clear if there is any motivation for them to change or improve their behavior or if there is justice for victims (Bobo & Thompson, 2006; Chandrasekharan et al., 2017; Cole, 1999; Schoenebeck et al., 2020). Schoenebeck et al. (2021) note that a public apology can be important parts of a justice procedure, although apologies may be harmful if delivered poorly (Battistella, 2014; Mingus, 2019). Offender lists are rated highly as a potential response to harassment (Schoenebeck et al., 2020), but are noted as a kind of public shaming, which has been generally discarded in legal and moral thought as a violation of human rights and dignity (Klonick, 2015; Nussbaum, 2009).

We also consider a remedy from Goldman (2021) that was not included in Schoenebeck et al.: labeling content. Labeling a violation of a community's rules can be an important part of educating a community and setting norms about what is appropriate behavior or not (Matias, 2019, n.d.). It can also be an important part of validating the experiences of harassment targets—knowing that a social media site labeled content as violating their guidelines can help targets feel heard and supported (Blackwell et al., 2017). Schoenebeck et al. find that participants are favorable toward content removal, user bans, and apologies (and are unfavorable toward other remedies like mediation that are described in their study). However, they do not assess whether preferences for remedies vary by online harassment type, which is important for selecting remedies appropriate to harassment contexts. Thus, our third research question asks the following:

RQ3.: What are the preferred remedies associated with different types of online harassment?

Prior work suggests that preferences for remedies to online harassment may vary by identity. For example, some marginalized groups may not prefer apologies as remedies because the apologies may be inauthentic (Schoenebeck et al., 2020). However, it is not known whether these preferences vary by type of harassment. Thus, our last research question is as follows:

RQ4. How do preferred remedies associated with types of online harassment vary by identity?

Table 1. Harassment Types Used in Study 1.

Harassment type	Survey prompt
Hateful Names	Called you hateful names on social media based on your identity
Doxxing*	Posted your home address on social media and threatened you with physical harm
Spread Rumors	Spread malicious rumors about you on social media
Share Sexual Photos*	Shared sexually explicit photos of you on social media
Insult/Disrespect	Insulted or disrespected you on social media
Send Unsolicited Photos	Sent you unsolicited sexually explicit photos on social media
Embarrass	Created a fake account pretending to be you on social media and posted embarrassing content
Damage Reputation*	Created a fake account pretending to be you on social media and posted content that damaged your reputation
Ban Account	Reported your account to a social media site that resulted in you being banned from the site
Harass Family Friends*	Harassed your friends and family on social media as a way to hurt you
Harass Employer	Harassed your employer on social media as a way to hurt you

Types with * were used in Studies 2 and 3.

Table 2. Remedies For Online Harassment Used In Study 3.

Remedy	Survey prompt
Content Removal	Removing the content from the site
Flag	Flagging the content as inappropriate
User Ban	Banning the person from the site
Public List	Adding the person to an online public list of offenders
Payment	Paying you and people who support you
Public Apology	Requiring a public apology from the person

Methods

We conducted three online surveys with US adult participants. Study 1 addresses RQs 1 and 2. Study 2 addresses H1a and H1b. Study 3 addresses RQs 3 and 4. We chose to conduct three separate surveys because the designs of Studies 2 and 3 relied on the results from Study 1, and because Study 2 used a between-subject design that would prime participants, confounding results of Study 3 if they were combined.

Survey Design and Analysis

Study 1 measured perceived harms associated with 11 harassment types (see Table 1). For each harassment type, we asked participants to respond to three questions to gauge the degree of perceived physical harm, sexual harm, and psychological harm associated with the type of harassment. Participants indicated their response on a 7-point scale ranging from "Not at all" to "Extremely." Participants also completed demographic questions. We report results from Study 1 below. We also used results from Study 1 to narrow the 11 types down to a subset of four types that were higher in harm for use in Studies 2 and 3. The first three—Doxxing, Share Sexual Photos, and Damage Reputation—were the highest three in harm. The next two—Send Unsolicited Photos and Embarrass—are similar to the first three, so we used the sixth highest—Harass Friends Family—as the fourth type for Studies 2 and 3.

Study 2 measured harm associated with group behaviors and frequency of harassment for four types of harassment. This was a between-subject study with two independent variables: group versus individual harassment, and repeated versus one-time harassment. Participants only saw one of the four possible conditions via random assignment and rated perceived physical, sexual, and psychological harms.

Study 3 measured remedy preferences. We conducted a mixed between-within subject study to evaluate preferred remedies associated with the four harassment types and the two groups and frequency variables. The within-subject conditions were the four harassment types and six possible remedies to the harassment (see Table 2). The between subjects were the group and frequency variables, similar to Study 2. We randomly assigned participants to one of four between-subject conditions. Participants were presented with the four harassment types, one per page, and answered questions about to what extent each of the six response options are desirable, fair, and just based on measures used in the work of Schoenebeck et al. (2020). The three dependent variables correlated highly with one another, so we averaged them together to create a single dependent variable of preference for the remedy ($\alpha = .91$). We conducted analyses using linear mixed models to account for the nested nature of our data, where each participant provided responses to multiple harassment types. We report 95% confidence intervals (CIs) in Studies 2 and 3.

Demographics: All three studies asked demographic questions. The demographics questions asked about age, gender identity, transgender identity, sexual orientation, race, education, employment, disability, income, political views, and location (city, state, and country). Studies 2 and 3 also about social media experiences (four questions) and social media use (one matrix question). The social media experiences questions asked how often participants had been harassed on social media, how often they had supported someone else, how often they had harassed others, and an open-ended question. The social media use question asked about frequency of use of nine major social media platforms on a 7-point scale of *I do not use it* to *All or most of the day*.

Participant Recruitments and Demographics

We recruited participants from Prolific.co, an online platform for recruiting participants for research studies. Study 1 was completed by 340 participants who received $3 as compensation for their time. The survey took less than 5 min to complete on average. Study 2 was completed by 615 participants who were compensated US$1 and the survey took approximately 1.5 min to complete. Study 3 was completed by 650 participants who received US$3 compensation and the survey took approximately 7.5 min to complete. The average compensation per hour based on survey average completion time was roughly US$36/hr (Study 1), US$40/hr (Study 2), and US$24/hr (Study 3). All participants were located in the United States. The studies were approved by our Institutional Review Board.

We aimed to sample participants from a range of identities and backgrounds based on gender, political views, race, disability, and sexual orientation (Duggan, 2017; Gray, 2012; Nakamura, 2013). We did this by conducting recruitment in batches where we could oversample some groups each time: for example, Study 1 was conducted in eight batches and we oversampled by gender diversity, racial diversity, political diversity, and so on. Demographics across the three studies were similar to each other: the median age in each of the studies ranged from 35 to 37 (SD=12–15). Participant genders across the studies were 49%–58% women; 35%–43% men; and 7%–8% non-binary. Around 5% were transgender. About 27%–31% identified as having a sexual orientation other than heterosexual (e.g., bisexual). About 34%–36% identified as people with a disability. Around 60%–70% were employed full- or part-time, and their median incomes were US$30,000–US$50,000. They were more liberal (45%–51%) than conservative (24%–33%) or moderate (22%–25%).

Results

Results are organized in two sections. The first section analyzes perceived harms associated with online harassment (RQs 1, 2; H1a, H1b) and how perceptions vary by identity (R2). The second section analyzes preferred remedies for online harassment (RQ3) and how preferences vary by identity (RQ4).

Perceived Harm

Because each participant rated multiple types of harassment, we performed linear mixed models to account for the nested nature of the data, with harm as the dependent variable and the types of harassment as the independent variable. Figure 1 illustrates individual harm levels by harassment type.

To address RQ1 (what are the perceived harms associated with online harassment), we first summed the ratings of sexual harm, physical harm, and psychological harm together to measure overall harm. Share Sexual Photos was the highest in overall harm (M=4.87, SD=1.35), followed by Doxxing (M=4.56, SD=1.36) and Damage Reputation (M=3.66, SD=1.28). A Tukey honestly significant difference (HSD) comparison of means test shows that all three were significantly different from each other and from the other types of harassment. The lowest types of harm were Insult/Disrespect (M=2.4, SD=1.16) and Ban Account (M=2.43, SD=1.14), which were not significantly different from each other but were significantly lower than all the other types.

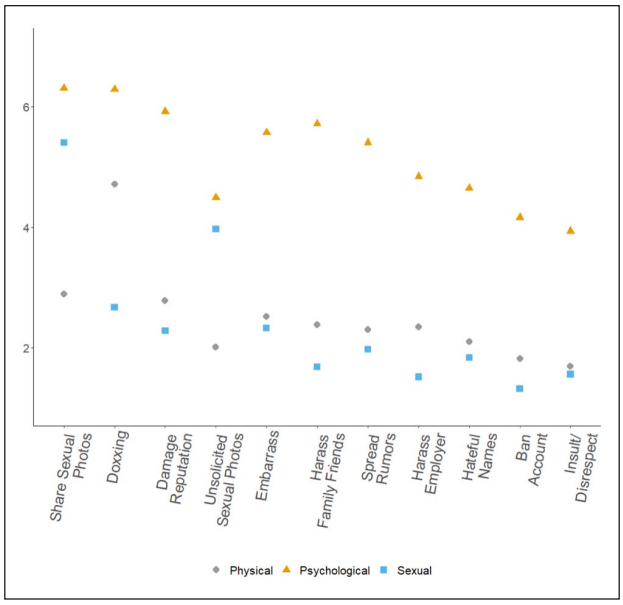

Figure 1. Perceived physical, sexual, and psychological harms by 11 harassment types, in descending order of average harm across three types. *Y*-axis is measure of harm with 1 = *not at all* and 7 = *extremely*.

We then examined individual levels of harm. Doxxing was rated higher in physical harm than other types (M=4.72, SD=2.03). Sharing Sexual Photos (M=5.41, SD=1.98) and Send Unsolicited Photos (M=4.49, SD=2.09) were highest in sexual harm, followed by Doxxing (M=2.67, SD=1.94). Doxxing (M=6.29, SD=1.2) and Share Sexual Photos (M=6.31, SD=1.31) were highest in psychological harm. On average, psychological harm was rated higher (i.e., more harmful) across the 11 harassment types (M=5.21, SD=1.87), while physical harm (M=2.5, SD=1.93) and sexual harm (M=2.41, SD=1.96) were rated lower.

To address H1a and H1b (perceptions of harm associated with frequency and group size), we turned to Study 2. Results show that repeated online harassment is perceived as higher in sexual harm (β=.243, p=.027, CI=0.289, .458) and psychological harm (β=.176, p=.047, CI=.003, .350) than one-time harassment, but not physical harm. There was no significant difference in perceptions of harm between harassment by one person versus a group of people.

To address RQ2 (how do perceptions of harm vary by identity), we investigate how perception of harm and preferred remedies varies by harassment type and demographics. We ran regression models with perception of harm or preference for remedy as the dependent variables. We used the stepAIC function in R to determine the model with the best (lowest) Akaike information criterion (AIC) score and present only this model. We ran 16 models to predict harm, across four harassment types and four harm measures (the three types of harm plus total or combined harm). Full model outputs are in given the Appendix. The most frequently significant predictor was age (13/16 models), where participants who are older report greater perceived harm of harassment. Gender was the second most frequently significant factor (12/16 models), with women perceiving greater harm associated with harassment compared to men (nonbinary gender was present in one model). Participants who were White perceived less sexual and physical harm to Damage Reputation and Harass Family Friends compared to participants who were not White. Prior experiences with harassment and social media use showed up as predictors in the models, but there were not consistent themes across models (see Appendix).

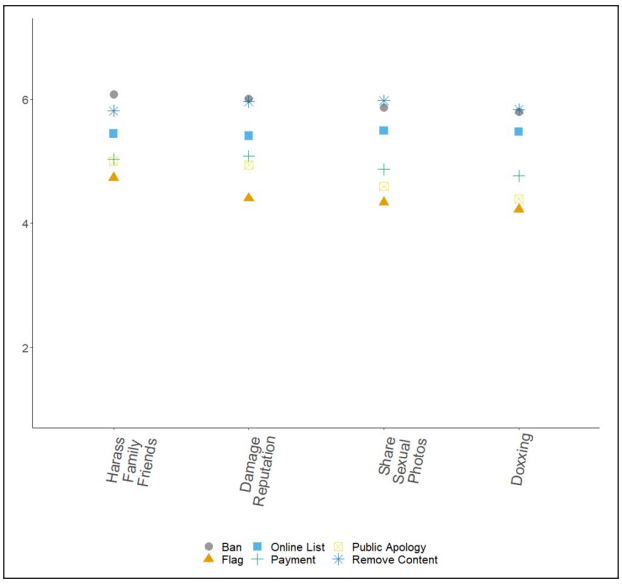

Figure 2. Preferences for six remedies by four harassment types, in descending order of average preference across six types. *Y*-axis is an averaged measure of preference from 1 (*lowest preference*) to 7 (*highest preference*).

Preferred Remedies

To examine RQ3 (what remedy preferences are associated with harassment types), we turn to Study 3. We first present an overview of remedy preferences and then analyze them by harassment type.

Overall, User Ban was the preferred remedy ($M=17.82$, $SD=4.14$), followed by Content Removal ($M=17.7$, $SD=4.15$), Public List ($M=16.38$, $SD=5.07$), Payment ($M=14.8$, $SD=5.46$), Public Apology ($M=14.18$, $SD=5.89$), and Flag ($M=13.29$, $SD=6.12$). A Tukey HSD comparison of means test shows that the difference between all types of responses was significant, except for between Content Removal and User Ban that were rated similarly favorably.

Preferences were higher for some remedies when the harassment type was Harass Family Friends or Damage Reputation but lower when the harassment was Share Sexual Photos or Doxxing (the harassment types that were highest in harm). This difference was statistically significant for Payment, User Ban, and Public Apology but not for the other remedies (see Figure 2). More generally, participants reported lower preferences for remedies in response to harassment types rated higher in physical harm ($\beta=-.121$, $p<.001$, CI=–.174, .069) and sexual harm ($\beta=-.046$, $p=.03$, CI=–.087, .004). There was no significant relationship ($\beta=.054$, $p=.256$, CI=.039, .146) between psychological harm and preferences for remedies.

For predicting satisfaction with remedies by identity, we ran 24 models across four harassment types and six type of responses. Gender (woman) was the most frequently significant predictor in 12 of 24 models. Specifically, women reported lower satisfaction than men with the payment remedy for Doxxing and Damage Reputation. However, women reported higher satisfaction with the responses of Removing Content (all four harassment types), Banning Harassers (three harassment types except Doxxing), and adding harassers to public Online Lists (three harassment types except Damage Reputation). Participants who were White were more favorable toward Removing Content for Damage Reputation and Harass Family Friends. Disability was in 5 of the 24 models; for all 5, people who had a disability were less favorable toward the remedy. For example, participants with a disability were less favorable toward Banning for the harassment types: Doxxing and Harass Family Friends.

Limitations and Considerations

Surveys have a number of known limitations. One major one for this study is that participants' expressed perceptions and preferences may not align with their actual perceptions and preferences. Without measuring real-time reactions to actual harassment experiences, our surveys measure expressed attitudes only. We decided to conduct three distinct surveys that helped simplify the study designs of each; however, this precludes us from analyzing patterns across the studies. Relying on a sample of prolific users will yield responses that are not representative of the general population; we mitigated sampling biases where possible with oversampling but not in a way that addressed biases associated with Prolific itself. Quantitative comparisons, like the ones we have used, should always be taken with a grain of salt. It is likely that some of the patterns might shift in another study or even rerunning the

same study. We tried to focus on high level patterns and themes in our reporting, especially of the regression analyses, to focus on more robust patterns. Finally, relying on a U.S. sample overlooks the majority of the world. To the extent we care about what US participants say, we should also care about what people in any other region of the world say.

Discussion

This section reflects on three prominent themes: the importance of disambiguating online harassment discussions into specific types of harassment, the limitations of existing remedies, and the need for victim-centered responses to repeated harassment.

Accounting for Harassment Types and Harm in Design and Policy

There were substantial differences in perceived harm based on harassment type, with share sexual photos rated highest (4.87 on a 7-point scale) and insult/disrespect lowest (2.4). Results also indicated that perceptions of harm differ by identity characteristics, including age, gender, and race. Of particular note is that the groups who perceived higher harm—women, older people, and people of color—are those who are underrepresented and excluded from most tech policy and from positions of power more generally. As many scholars have noted (e.g., Benjamin, 2019; Noble, 2018), if companies want to improve their responses to treatments of marginalized groups, they need to prioritize the voices of those who experience harm in their policy making and organizations. Current social media governance practices focus on the content that was posted and the user who posted it, relying on punitive models used in criminal legal systems (Schoenebeck & Blackwell, 2021). Though these systems are needed in online contexts, an additional focus on harassment experiences could draw from trauma-informed practice, care frameworks, affirmative consent (Chen et al., 2022; Im et al., 2021; Tseng et al., 2022), or other principles, all of which could be designed into social media experiences.

Currently, social media platforms may have categories of content violations but they tend to be developed via top-down policy rather than from bottom-up user studies (Fisher, 2018). Understanding how users' experiences of harm vary by harassment type is important for developing responses that appropriately address the harm they experience. Psychological harm was more strongly associated with all harassment types, whereas physical and sexual harms were only strongly associated with some types. Currently, the law recognizes physical harms (e.g., bodily harm) and sexual harm (e.g., rape) but is less consistent in its recognition of psychological harm or in other types of harm, which are not typically cognizable. Citron and Solove (2022) note that the requirement of harm has impeded the enforcement of law (in the context of privacy) and propose that developing a taxonomy of harms is important for courts to tackle them in any meaningful way. While this study was not developed with an explicit eye toward regulatory implications, developing taxonomies of harm could be important for platforms to better understand how their policy and design choices may be meeting or overlooking users' needs.

Expanding Remedies for Reputational Harm

Results show that participants rated remedy preferences as favorable for all harassment types and remedy types. That is, they prefer any remedy over not having it. On average, they preferred Content Removal, User Ban, and Public Lists, which is consistent with prior work, though Public Apology was rated less positively in our study than in a prior study (Schoenebeck et al., 2020). However, prior work has not disentangled remedy preferences by harassment type.

Remedy preferences for Payment, User Banning, and Apology were higher for two harassment types—Harass Family Friends and Damage Reputation. Although reputational harm has been studied in the context of businesses (Eccles et al., 2007) and in the context of reputation-based systems like Yelp (Reagle, 2015), how to recover from personal reputational harm, especially outside of legal contexts (e.g., defamation) and on unbounded social media platforms, is not well understood. We included relational and reputational types of harassment because they are not typically protected by law or policy even though they can contribute to severe, negative experiences online. These types of relational and reputational harm are also not well-studied in online harassment literature, which has tended to focus on harassment experienced by the individual (e.g., hate speech, sexual harassment). Reputational harm can be especially important in cultures and regions where honor is a prominent value; in those cases, harassment victims can be shamed or even murdered for a perceived transgression online (Maher, 2020).

Remedy Preferences May not Correspond with Severity of Harassment

Remedy preferences were lower for Payment, User Ban, and Public Apology when the harassment type was Share Sexual Photos or Doxxing. These two harassment types are higher in overall harm, suggesting that there may be a bimodal (or some other) relationship rather than a linear relationship between harm and remedy—that is, remedy preferences may increase as harm increases, but after a certain level they may decrease again. We interpret this possible relationship from our data as an indicator that even highly rated remedies may fail to effectively address and remediate harm to users. This may be because such approaches overlook or diminish the harms associated with harassment, or because users who are banned can often create new accounts or engaged in networked harassment via other users (Marwick, 2021).

Both non-consensual sharing of sexual photos and doxxing can be devastating experiences for victims, often requiring substantial legal recourse, financial resources, appeals to companies, and support from personal networks. Repeated harassment can invoke retraumatization, or secondary trauma (Bloom, 1999; Harms, 2015; Levenson, 2017), where a person who has been harassed repeatedly may experience ongoing retraumatization based on those experiences. The remedies presented in this study likely are too insignificant and individual to address those types of harm, which are rooted in deep-seated societal injustices. After extensive advocacy (Citron, 2022; Goldberg, 2019), non-consensual sharing of intimate content is now recognized by courts in most states in the United States, though access to that legal process requires resources and know-how. There should be more attention to how to design platforms that deter sharing of non-consensual sharing of content and doxxing, and that support victims who are targeted by those campaigns. For example, systems can be designed to help users protect themselves from the most severe types of harassment, including repeated harassment, through elevated blocking mechanisms, facilitation of support seeking, and automated reporting systems.

Conclusion

Currently, social media sites do little to distinguish between types of harassment and their associated harms—most sites rely on a small set of remedies to attend to a wide range of behaviors. Furthermore, those remedies tend to focus on sanctions to people who caused the harassment, while overlooking how people who experience harassment might want to experience redress. In addition, some harassment types are more harmful than others, such as doxxing and non-sharing of sexual photos, and remedies for those experiences may fall short of actually remediating harm. Our article sheds light on how individualistic and varied experience of online harassment can be when it comes to perceptions of physical, sexual, and psychological harms, as well as the relationship between the type of harassment and the appropriate remedy for them. Our findings demonstrate how the victim's identity, such as their minority status, can further complicate these perceptions of harms and preferences for appropriate remedies. Social media platforms should consider harm when deciding how to respond to harassment, and should differentiate between harassment experiences and associated harms to determine how to sanction offenders and support victims.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This material is based on work supported by the National Science Foundation under Grant Nos 1763297 and 1552503. They thank members of the Social Media Research Lab for their feedback.

ORCID iDs

Sarita Schoenebeck https://orcid.org/0000-0002-8688-1595
Penny Triệu https://orcid.org/0000-0001-8394-803X

Supplemental Material

Supplemental material for this article is available online.

References

Alhaboby, Z. A., al-Khateeb, H. M., Barnes, J., & Short, E. (2016). 'The language is disgusting and they refer to my disability': The cyber harassment of disabled people. *Disability & Society*, *31*(8), 1138–1143. https://doi.org/10.1080/09687599.2016.1235313

Aziz, Z. A. (2015). *Online violence against women and engendering digital equality*. Office of the High Commissioner for Human Rights.

Bailey, M. (2021). *Misogynoir transformed*. NYU Press. https://nyupress.org/9781479865109/misogynoir-transformed

Battistella, E. L. (2014). *Sorry about that: The language of public apology*. Oxford University Press.

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new Jim code*. John Wiley & Sons.

Blackwell, L., Chen, T., Schoenebeck, S., & Lampe, C. (2018). When online harassment is perceived to be justified. *International AAA Conference on Web and Social Media (ICWSM 2018)*. http://www.lindsayblackwell.net/wp-content/uploads/2018/04/When-Online-Harassment-is-Perceived-as-Justified-ICWSM18.pdf

Blackwell, L., Dimond, J., Schoenebeck, S., & Lampe, C. (2017). Classification and its consequences for online harassment: Design insights from HeartMob. *Proceedings of the ACM on Human-Computer Interaction*, *1*(2), 19.

Bloom, S. L. (1999). Trauma theory abbreviated. *CommunityWorks*, *14*. https://strengthcounselling.ca/wp-content/uploads/2018/05/trauma-theory-abbreviated.pdf

Bobo, L. D., & Thompson, V. (2006). Unfair by design: The war on drugs, race, and the legitimacy of the criminal justice system. *Social Research: An International Quarterly*, *73*(2), 445–472.

Bradford, B., Grisel, F., Meares, T. L., Owens, E., Pineda, B. L., Shapiro, J., Tyler, T. R., & Peterman, D. E. (2019). *Report of the Facebook data transparency advisory group*. Yale Justice Collaboratory.

Bridy, A. (2018). Remediating social media: A layer-conscious approach. *Boston University Journal of Science and Technology Law*, *24*, 193.

Center for Disease Control. (2021, February 23). *Intimate partner violence prevention*. https://www.cdc.gov/violenceprevention/intimatepartnerviolence/index.html

Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., & Gilbert, E. (2017). You can't stay here: The efficacy of Reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction*, *31*, 22. https://doi.org/10.1145/3134666

Chandrasekharan, E., Samory, M., Jhaver, S., Charvat, H., Bruckman, A., Lampe, C., Eisenstein, J., & Gilbert, E. (2018). The internet's hidden rules: An empirical study of Reddit norm violations at micro, Meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 32:1–32:25. https://doi.org/10.1145/3274301

Chen, J. X., McDonald, A., Zou, Y., Tseng, E., Roundy, K. A., Tamersoy, A., Schaub, F., Ristenpart, T., & Dell, N. (2022). Trauma-informed computing: Towards safer technology experiences for all. In *Proceedings of the 2022 CHI conference on human factors in computing systems* (pp. 1–20). https://nixdell.com/papers/chi22-trauma-informed-computing.pdf

Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2017). Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing* (pp. 1217–1230). https://www.cs.cornell.edu/~cristian/Anyone_Can_Become_a_Troll_files/anyone_can_become_a_troll.pdf

Cheng, J., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2015, April 21). *Antisocial behavior in online discussion communities*. Ninth International AAAI Conference on Web and Social Media. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10469

Citron, D. K. (2017). Extremist speech, compelled conformity, and censorship creep. *Notre Dame Law Review*, *93*, 1035.

Citron, D. K. (2022). *The fight for privacy: Protecting dignity, identity and love in the digital age*. Penguin. https://www.penguin.co.uk/books/446475/the-fight-for-privacy-by-citron-danielle-keats/9781784744847

Citron, D. K., & Solove, D. J. (2022). Privacy harms. *Boston University Law Review*, *102*, 793.

Cole, D. (1999). *No equal justice: Race and class in the American criminal justice system*. New Press. https://www.ncjrs.gov/App/abstractdb/AbstractDBDetails.aspx?id=179184

Douek, E. (2020). *The limits of international law in content moderation* (SSRN Scholarly Paper ID 3709566). Social Science Research Network. https://doi.org/10.2139/ssrn.3709566

Dragiewicz, M., Burgess, J., Matamoros-Fernández, A., Salter, M., Suzor, N. P., Woodlock, D., & Harris, B. (2018). Technology facilitated coercive control: Domestic violence and the competing roles of digital media platforms. *Feminist Media Studies*, *18*(4), 609–625. https://doi.org/10.1080/14680777.2018.1447341

Duggan, M. (2017). Online harassment 2017. *Pew Research Center*. http://www.pewinternet.org/2017/07/11/online-harassment-2017/

Eccles, R. G., Newquist, S. C., & Schatz, R. (2007, February 1). Reputation and its risks. *Harvard Business Review*. https://hbr.org/2007/02/reputation-and-its-risks

Fisher, M. (2018, December 27). Inside Facebook's Secret Rulebook for Global Political Speech. *The New York Times*. https://www.nytimes.com/2018/12/27/world/facebook-moderators.html

Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Goldberg, C. (2019). *Nobody's victim: Fighting psychos, stalkers, pervs, and trolls*. Penguin.

Goldman, E. (2021). Content moderation remedies. *Michigan Technology Law Review*, *28*. https://doi.org/10.2139/ssrn.3810580

Gray, K. L. (2012). Intersecting oppressions and online communities. *Information, Communication & Society*, *15*(3), 411–428. https://doi.org/10.1080/1369118X.2011.642401

Gray, K. L. (2020). Black gamers' resistance. In *Race and media: Critical approaches* (p. 241). NYU Press. https://www.degruyter.com/document/doi/10.18574/nyu/9781479823222.003.0023/pdf

Harms, L. (2015). *Understanding trauma and resilience*. Macmillan International Higher Education.

Heise, L. (1994). Gender-based abuse: The global epidemic. *Cadernos De Saúde Pública*, *10*, S135–S145. https://doi.org/10.1590/S0102-311X1994000500009

Hinduja, S., & Patchin, J. W. (2010). Bullying, cyberbullying, and suicide. *Archives of Suicide Research*, *14*(3), 206–221. https://doi.org/10.1080/13811118.2010.494133

Im, J., Dimond, J., Berton, M., Lee, U., Mustelier, K., Ackerman, M. S., & Gilbert, E. (2021). Yes: Affirmative consent as a theoretical framework for understanding and imagining social platforms. *Proceedings of the 2021 CHI conference on human factors in computing systems*. https://imjane.net/papers/chi2021-affirmative-consent.pdf

Jhaver, S., Bruckman, A., & Gilbert, E. (2019). Does transparency in moderation really matter? User behavior after content removal explanations on Reddit. https://dl.acm.org/doi/pdf/10.1145/3359252#:~:text=This%20lack%20of%20transparency%20of,why%20their%20content%20was%20removed

Jiang, J. A., Scheuerman, M. K., Fiesler, C., & Brubaker, J. R. (2021). Understanding international perceptions of the severity of harmful content online. *PLOS ONE*, *16*(8), e0256762. https://doi.org/10.1371/journal.pone.0256762

Kaye, D. (2019). *Speech police: The global struggle to govern the internet*. Columbia Global Reports.

Kazerooni, F., Taylor, S. H., Bazarova, N. N., & Whitlock, J. (2018). Cyberbullying bystander intervention: The number of offenders and retweeting predict likelihood of helping a cyberbullying victim. *Journal of Computer-mediated Communication*, *23*(3), 146–162. https://doi.org/10.1093/jcmc/zmy005

Klonick, K. (2015). Re-shaming the debate: Social norms, shame, and regulation in an internet age. *Maryland Law Review*, *75*, 1029.

Kowalski, R. M., Giumetti, G. W., Schroeder, A. N., & Lattanner, M. R. (2014). Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth. *Psychological Bulletin*, *140*(4), 1073–1137. https://doi.org/10.1037/a0035618

Langos, C. (2015). Cyberbullying: The shades of harm. *Psychiatry, Psychology and Law*, *22*(1), 106–123. https://doi.org/10.1080/13218719.2014.919643

Lenhart, A., Ybarra, M., Zickuhr, K., & Price-Feeney, M. (2016, November 21). *Online harassment, digital abuse, and cyberstalking in America*. Data & Society, Data & Society Research Institute. https://datasociety.net/library/online-harassment-digital-abuse-cyberstalking/

Levenson, J. (2017). Trauma-informed social work practice. *Social Work*, *62*(2), 105–113.

Lirri, E. (2015). *The challenge of tackling online violence against women in Africa*. https://www.opennetafrica.org/the-challenge-of-tackling-online-violence-against-women-in-africa/

Maher, S. (2020). *A woman like her: The story behind the honor killing of a social media star*. Melville House.

Marwick, A. (2021). Morally Motivated Networked Harassment as Normative Reinforcement. *Social Media+Society*, *7*(2), 20563051211021376. https://doi.org/10.1177/20563051211021378

Marwick, A., Blackwell, L., & Katherine, L. (2019). *Best practices for conducting risky research and protecting yourself from online harassment (Data & society guide)*. Data & Society Research Institute.

Massanari, A. (2017). #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*, *19*(3), 329–346. https://doi.org/10.1177/1461444815608807

Matias, J. N. (2019). Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences*, *116*(20), 9785–9789. https://doi.org/10.1073/pnas.1813486116

Matias, J. N. (n.d.). Posting Rules in Online Discussions Prevents Problems & Increases Participation. *CivilServant*. http://civilservant.io/r_science_sticky_coments_1.html

Mingus, M. (2019, December 18). The four parts of accountability: How to give a genuine apology part 1. *Leaving Evidence*. https://leavingevidence.wordpress.com/2019/12/18/how-to-give-a-good-apology-part-1-the-four-parts-of-accountability/

Nakamura, L. (2013). *Cybertypes: Race, ethnicity, and identity on the internet*. Routledge.

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. NYU Press.

Nussbaum, M. C. (2009). *Hiding from humanity: Disgust, shame, and the law*. Princeton University Press.

Online Violence & Internet Harassment of women. (2015). Global Fund for Women. https://www.globalfundforwomen.org/online-violence-just-because-its-virtual-doesnt-make-it-any-less-real/

Pater, J. A., Kim, M. K., Mynatt, E. D., & Fiesler, C. (2016). Characterizations of online harassment: Comparing policies across social media platforms. In *Proceedings of the 19th international conference on supporting group work* (pp. 369–374). https://dl.acm.org/doi/10.1145/2957276.2957297

Perez, S. (n.d.). Twitter adds more anti-abuse measures focused on banning accounts, silencing bullying. *TechCrunch*. http://social.techcrunch.com/2017/03/01/twitter-adds-more-anti-abuse-measures-focused-on-banning-accounts-silencing-bullying/

Pskowski, M. (2017). Mexican women stand up to cyberattacks and vicious digital violence. *The World*. https://www.pri.org/stories/2017-08-24/mexican-women-stand-cyberattacks-and-vicious-digital-violence

Reagle, J. M. (2015). *Reading the comments: Likers, haters, and manipulators at the bottom of the web*. MIT Press.

Reed, L. A., Tolman, R. M., & Ward, L. M. (2017). Gender matters: Experiences and consequences of digital dating abuse victimization in adolescent dating relationships. *Journal of Adolescence*, *59*, 79–89. https://doi.org/10.1016/j.adolescence.2017.05.015

Roberts, S. T. (2019). *Behind the screen: Content moderation in the shadows of social media*. Yale University Press.

Rocha-Silva, T., Nogueira, C., & Rodrigues, L. (2021). Intimate abuse through technology: A systematic review of scientific Constructs and behavioral dimensions. *Computers in Human Behavior*, *122*, 106861. https://doi.org/10.1016/j.chb.2021.106861

Scheuerman, M. K., Jiang, J. A., Fiesler, C., & Brubaker, J. R. (2021). A framework of severity for harmful content online. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 368:1–368:33. https://doi.org/10.1145/3479512

Schoenebeck, S., & Blackwell, L. (2021). Reimagining social media governance: Harm, accountability, and repair. *Yale Journal of Law and Technology*, *13*. https://law.yale.edu/sites/default/files/area/center/justice/reimagining_social_media_governance_harm_accountability_and_repair.pdf

Schoenebeck, S., Haimson, O. L., & Nakamura, L. (2020). Drawing from justice theories to support targets of online harassment. *New Media & Society*, *23*, 1278–1300. https://doi.org/10.1177/1461444820913122

Schoenebeck, S., Scott, C., Hurley, E., Chang, T., & Selkie, E. (2021). Youth trust in social media companies' responses to online harassment. PACM HCI. http://yardi.people.si.umich.edu/pubs/Schoenebeck_YouthTrust2021.pdf

Slonje, R., & Smith, P. K. (2008). Cyberbullying: Another main type of bullying? *Scandinavian Journal of Psychology*, *49*(2), 147–154. https://doi.org/10.1111/j.1467-9450.2007.00611.x

SoK: Hate, harassment, and the changing landscape of online abuse. (2021). In K. Thomas, D. Akhawe, M. Bailey, D. Boneh, E. Bursztein, S. Consolvo, N. Dell, Z. Durumeric, P. G. Kelley, D. Kumar, D. McCoy, S. Meiklejohn, T. Ristenpart, & G. Stringhini (Eds.), *Proceedings of the IEEE Symposium on Security and Privacy*. https://research.google/pubs/pub49786/

Solove, D. J., & Citron, D. K. (2017). Risk and anxiety: A theory of data-breach harms. *Texas Law Review*, *96*(4), 737–786.

Stevenson, A. (2018, November 6). Facebook admits it was used to incite violence in Myanmar. *The New York Times*. https://www.nytimes.com/2018/11/06/technology/myanmar-facebook.html

Stewart, A., Schuschke, J., & Tynes, B. (2019). Online racism: Adjustment and protective factors among adolescents of color. In H. E. Fitzgerald, D. J. Johnson, D. B. Qin, F. A. Villarruel, & J. Norder (Eds.), *Handbook of children and prejudice: Integrating research, practice, and policy* (pp. 501–513). Springer. https://doi.org/10.1007/978-3-030-12228-7_28

Stroud, S. (2016). "Be a bully to beat a bully": Twitter ethics, online identity, and the culture of quick revenge. In A. Davisson & P. Booth (Eds.), *Controversies in digital ethics* (pp. 264–278). Bloomsbury.

Tseng, E., Bellini, R., McDonald, N., Danos, M., Greenstadt, R., McCoy, D., Dell, N., & Ristenpart, T. (2020). The tools and tactics used in intimate partner surveillance: An analysis of online infidelity forums. *29th USENIX Security Symposium USENIX Security 20*, (USENIX Security 20), 1893–1909. http://nixdell.com/papers/Tseng-2020-USENIX-Sec_Tools-Tactics-Intimate-Partner-Surveillance.pdf

Tseng, E., Sabet, M., Bellini, R., Sodhi, H. K., Ristenpart, T., & Dell, N. (2022). Care infrastructures for digital security in intimate partner violence. CHI Conference on Human Factors in Computing Systems. https://dl.acm.org/doi/pdf/10.1145/3491102.3502038

UNESCO. (2021, March 8). *UNESCO calls to end online violence against women journalists in 8 March campaign*. https://en.unesco.org/news/unesco-calls-end-online-violence-against-women-journalists-8-march-campaign

Usher, N., Holcomb, J., & Littman, J. (2018). Twitter makes it worse: Political journalists, gendered echo chambers, and the amplification of gender bias. *The International*

Journal of Press/politics, *23*(3), 324–344. https://doi.org/10.1177/1940161218781254

World Health Organization. (2021). *Violence against women*. https://www.who.int/news-room/fact-sheets/detail/violence-against-women

York, J. C. (2021). *Silicon values: The future of free speech under surveillance capitalism*. Verso Books.

Author Biographies

Sarita Schoenebeck (PhD Georgia Institute of Technology) is an Associate Professor in the School of Information at the University of Michigan. Her research interests include equity and justice online, privacy, and technology policy.

Cliff Lampe (PhD/MSI University of Michigan) is a Professor at the School of Information in the University of Michigan. His research interests include the effects of social media use, interface elements as sociotechnical architecture, and civic technology.

Penny Triệu received a PhD in Information Science from the University of Michigan. Her research interests include computer-mediated communication, and the interpersonal and psychological implications of social media.