Online Harassment in Majority Contexts: Examining Harms and Remedies across Countries

Sarita Schoenebeck
University of Michigan
USA

Amna Batool
University of Michigan
USA

Giang Do
University of Michigan
USA

Sylvia Darling
University of Michigan
USA

Gabriel Grill
University of Michigan
USA

Daricia Wilkinson
Clemson University
USA

Mehtab Khan
Yale University
USA

Kentaro Toyama
University of Michigan
USA

Louise Ashwell
University of Michigan
USA

ABSTRACT

Online harassment is a global problem. This article examines perceptions of harm and preferences for remedies associated with online harassment with nearly 4000 participants in 14 countries around the world. The countries in this work reflect a range of identities and values, with a focus on those outside of North American and European contexts. Results show that perceptions of harm are higher among participants from all countries studied compared to the United States. Non-consensual sharing of sexual photos is consistently rated as harmful in all countries, while insults and rumors are perceived as more harmful in non-U.S. countries, especially harm to family reputation. Lower trust in other people and lower trust in sense of safety in one's neighborhood correlate with increased perceptions of harm of online harassment. In terms of remedies, participants in most countries prefer monetary compensation, apologies, and publicly revealing offender's identities compared to the U.S. Social media platform design and policy must consider regional values and norms, which may depart from U.S. centric-approaches.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing theory, concepts and paradigms**; **Empirical studies in collaborative and social computing**.

KEYWORDS

online harassment; online abuse; trust; courts; majority world; Non-Western; global south; social media; online governance

1 INTRODUCTION

Online harassment is pervasive in regions around the world. Users post hate speech that demeans and degrades people based on their gender, race, sexual identity, or position in society [11, 60]; users post insults and spread rumors, disproportionately harming those with fewer resources in society to cope with or respond to the attacks [60, 64, 108]; and users share private, sensitive content, like home addresses or sexual images, without the consent of those whose information is being shared [38]. These behaviors introduce multiple types of harm with varied levels of severity, ranging from minor nuisances to psychological harm to economic precarity to life threats [55, 85, 87]. Gaining a global understanding of online harassment is important for designing online experiences that meet the needs of diverse, varied global experiences.

Social media platforms have struggled to govern online harassment, relying on human and algorithmic moderation systems that cannot easily adjudicate content that is as varied as the human population that creates it [39, 81]. Platforms maintain community guidelines that dictate what type of content is allowed or not allowed and then use the combination of human and automated pipelines to identify and address violations [37, 81]. However, identifying and categorizing what type of content is harmful or not is difficult for both humans and algorithms to do effectively and consistently. These challenges are magnified in multilingual environments where people may be trying to assess content in different languages or cultural contexts than they are familiar with, while algorithms are inadequately developed to work across these languages and contexts [43, 110].

Investigations of harms associated with online harassment have been given disproportionate attention in U.S. contexts. Most prominent technology companies are centered in the U.S., employing U.S. workers in executive positions and centering U.S. laws, norms, corporations, and people [13, 110]. Scholars have called attention to this problem, pointing out how experiences differ for people and communities globally (e.g. [85, 94, 110]). For example, a study of

199 South Asian women shows that they refrain from reporting abuse because platforms rarely have the contextual knowledge to understand local experiences [85]. Across countries, social media users have expressed distrust in platforms' ability to govern behavior effectively, especially systems that are vague, complicated, and U.S.- and European-centric [11, 22, 85].

Governing social media across the majority of the world requires understanding how to design platforms with policies and values that are aligned with the communities who use them. Towards that goal, this article examines perceptions of harm and preferences for remedies associated with online harassment via a survey conducted in 14 countries[1] around the world, selected for their diversity in location, culture, and economies. Results from this study shed light on similarities and differences in attitudes about harms and remedies in countries around the world. This work also demonstrates the complexities of measuring and making sense of these differences, which cannot be explained by a single factor and should not be assumed to be stable over time. This article advances scholarship on online harassment in majority contexts, and seeks to expand understandings about how to design platforms that meet the needs of the communities that use them.

2 IMPACTS OF ONLINE HARASSMENT

Online harassment is an umbrella term that encompasses myriad types of online behaviors including insults, hate speech, slurs, threats, doxxing, and non-consensual image sharing, among others. A rich body of literature has described characteristics of online harassment including what it is, who experiences it, and how platforms try to address it (e.g. [17, 29, 52, 65, 87, 95]). Microsoft's Digital Civility surveys and Google's state of abuse, hate, and harassment surveys indicate how harassment is experienced globally [68, 95]. Harassment can be especially severe when it is networked and coordinated, where groups of people threaten one or many other people's safety and wellbeing [64]. Other types of harassment are especially pernicious in-the-moment, such as reporting "crimes" so that law enforcement agencies investigate a home [10] or sharing a person's home address online with the intent of encouraging mobs of people to threaten that person at their home. Across types of harm, marginalized groups experienced disproportionate harm associated with harassment online, including racial minorities, religious minorities, caste minorities, sexual and gender minorities, and people who have been incarcerated [16, 32, 62, 76, 78, 102].

Sometimes users post malicious content intended to bypass community guidelines which are difficult to algorithmically detect [27, 99]. This makes it relatively easy to deceive automatic detection models by subtly modifying an otherwise highly toxic phrase so that the detection model assigns it a significantly lower toxicity score [48]. In addition, due to limited training on non-normative behavior, these automatic detection and classification tools can exacerbate existing structural inequities [48]. For instance, Facebook's removal of a photograph of two men kissing after flagging it as "graphic sexual content" highlighted the lack of inclusivity of non-dominant behavior in their automatic detection tools [49].

This valorization of certain viewpoints highlights that power resides among those who create these labels by embedding their own values and worldviews (mostly U.S.-centric) to classify particular behaviors as appropriate or inappropriate [11, 48].

The effects of harassment vary by experience and individuals but might include anxiety, stress, fear, humiliation, self-blame, anger, and illness. There is not yet a standard framework for measuring harms associated with online harassment, which can include physical harm, sexual harm, psychological harm, financial harm, reproductive harm, and relational harm [35]. These can manifest in myriad ways: online harassment can cause changes to technology use or privacy behaviors, increased safety and privacy concerns, and disruptions of work, sleep, and personal responsibilities [30, 40, 75]. Other consequences can include public shame and humiliation, an inability to find new romantic partners, mental health effects such as depression and anxiety, job loss or problems securing new employment, offline harassment and stalking, numerous mental health issues, such as post-traumatic stress disorder (PTSD), depression, anxiety, self-blame, self-harm, trust issues, low self-esteem, confidence, and loss of control [8, 9, 31, 73, 78, 84, 102]. These effects can be experienced for long periods of time due in part to the persistence and searchability of content [38]. Targets often choose to temporarily or permanently abstain from social media sites, despite the resulting isolation from information resources and support networks [38, 60].

Microsoft's Digital Civility Index, a yearly survey of participants in over 20 countries, indicates that men are more confident than women in managing online risks [68]. Sexual images of women and girls are disproportionately created, sent, and redistributed without consent which can severely impact women's lives [8, 9, 14, 28, 31]. In a study of unsolicited nude images and its affect on user engagement [45, 91], victims reported being bombarded with unwelcome explicit imagery and faced further insults when they attempted to reduce interaction. A survey by Maple et al. with 353 participants from the United Kingdom (68% of respondents were women) listed damage to their reputation as the primary fear of victims of cyberharassment [63].

The consequences of gendered and reputational harm can be devastating. In South Korea, celebrities Hara Goo and Sulli (Jin-ri Choi) died by suicide, which many attributed to the large-scale cyberbullying, sexual harassment, and gender violence they experienced online [46]. A social media Pakistani celebrity was murdered by her brother, who perceived her social media presence as a blemish on the family's honor [80]. Two girls and their mother were allegedly gunned down by a stepson and his friends over the non-consensual filming and sharing of a video of the girls enjoying rain among family [23]. Many of these harms are ignited and fueled by victim-blaming, where society places the responsibility solely on women and other marginalized groups to avoid being assaulted [18, 78, 102]. This blaming is also perpetuated digitally; for instance, a review of qualitative studies on non-consensual sharing highlighted that women are perceived as responsible if their images are shared because they voluntarily posed for and sent these images in the first place [102].

---

[1]Data was collected from 13 countries plus a collection of Caribbean countries. We use the term "country" throughout for readability.

3 CHALLENGES IN GOVERNING ONLINE HARASSMENT

Most social media sites have reporting systems aimed at flagging inappropriate content or behavior online [22]. Though platform policies do not explicitly define what constitutes online harassment [74], platforms have highlighted several activities and behaviors in their community guidelines including abuse, bullying, defaming, impersonation, stalking, and threats [54, 74]. Content that is reported goes into a processing pipeline where human workers evaluate the content and determine whether it violates community guidelines or not [81]. If it does, they may take it down and sanction the user who posted it, with sanctions ranging in severity from warnings to suspensions to permanent bans [39, 89]. Platforms use machine learning to automatically classify and filter out malicious content, abusive language, and offensive behaviors [17, 106, 109]. These range from adding contextual and semantic features in detection tools to generating computational models using preexisting data from online communities to using these machine learning models to assign "toxicity scores" [17, 106]. Though harassment detection approaches have improved dramatically, fundamental limitations remain [11], including false positives and true negatives, where content is taken down that should have stayed up and vice versa [44, 87].

Many of these problems are deeply embedded in algorithmic systems, which can reinforce Western tropes, such as associating the word "Muslim" with terrorists [4]. Algorithms to detect problematic content also perform substantially worse in non-English languages, perpetuating inequalities rather than remediating them [24]. Dominant voices can overrule automatically detected flagged content through situated judgments [22]. For instance, a widely distributed video of Neda, an Iranian woman caught up in street protests and shot by military police in 2009, was heavily flagged as violating YouTube's community guidelines for graphic violence, but YouTube justified leaving it up because the video was newsworthy [103].

Platform policies are written in complex terms that are inaccessible to many social media users, which makes it difficult for them to seek validation of their online harassment experiences [34]. Further, platform operators do not specify which prohibited activities are associated with which responses [74]. When combined with the punitive nature of sanctions, online governance systems may be confusing and ineffective at remediating user behavior, while overlooking the harms faced by victims of the behavior [89]. One alternative that has been proposed more recently is a focus on rehabilitation and reparation in the form of apologies, restitution, mediation, or validation of experiences [11, 89, 107]. Implementing responses to online harassment requires that users trust platforms' ability to select and implement that response [104]; however, public trust in technology companies has decreased in recent years, and there is also distrust of social media platforms' ability to effectively govern online behavior [2, 11, 70, 90]. 84% of social media users in the U.S. believe that it is the platform's responsibility to protect them from social media harassment [59], yet Lenhart et al.'s survey suggests that only 27% of victims reported harassing activities on these platforms [7]. A different survey by Wolak et al. with 1631 victims of sextortion found that 79% of victims did not report their situation to social media sites because they did not think it would be helpful to report [105]. Their participants indicated that platform reporting might be helpful only when victims are connected to perpetrators exclusively online which might be addressable through in-app reporting [105]. Sambasivan et al.'s study with 199 South Asian women revealed that participants refrain from reporting through platforms due to platforms' limited contextual understanding of victims' regional issues, which is further slowed by the platforms' requirements to fill out lengthy forms providing detailed contexts [85]. Musgrave et al. find that U.S. Black women and femmes do not report gendered and racist harassment because they do not believe reporting will help them [70].

Wolak et al. also found that only 16% of victims of sextortion reported their incidents to the police [105]. Many of those who reported to police described having a negative reporting experience, which deterred them from pursuing criminal charges against offenders [105]. Such experiences include police arguing for the inadequacy of proof to file complaints, that sextortion is a non-offensive act, lack of jurisdiction to take actions, and being generally rude, insensitive, and mocking [105]. Sambasivan et al. also reported that only a few of their nearly 200 participants reported abusive behaviors to police because they perceived law enforcement officers to have low technical literacy, to be likely to shame women, or to be abusers themseves. [85]. When abusers are persistent, even reporting typically does not address the ongoing harassment [38, 64].

Sara Ahmed introduces the concept "strategic inefficiency" to explain how institutions slow down complaint procedures that can then deter complaints from constituents [5]. The lack of formal reporting channels leads users to be largely self-reliant for mitigating and avoiding abuse. Techniques they use range from preventative strategies like limiting content, modifying privacy settings, self-censorship, using anonymous and gender-neutral identities, using humor, avoiding communication with others, ignoring abuse, confronting abusers, avoiding location sharing, deleting accounts, blocklists, changing contact information, changing passwords, using multiple emails accounts for different purposes, creating a new social media profile under a different name, blocking or unfriend someone and untagging themselves from photos [7, 21, 26, 36, 53, 66, 93, 99, 100, 105]. Whether reporting to companies or police, these approaches all put the burden of addressing harassment on the victims. If we want to better govern online behavior globally, we need to better understand what harms users experience and how platforms and policies can systematically better support them after those harms.

4 STUDY DESIGN

We conducted a cross-country online survey in 14 countries (13 countries plus multiple Caribbean countries). We aimed for a minimum of 250 respondents in each country which considered our desire for age variance and gender representation among men and women but without the higher sample size needed for representative samples or subgroup analyses. The survey focused on online harassment harms and remedies and included questions about demographics, personal values, societal issues, social media habits, and online harassment. This paper complements a prior paper from the same project that focused on gender [50]; this paper focuses

on country level differences though it also engages with gender as part of the narrative.

We iteratively designed the survey as a research team, discussing and revising questions over multiple months. When we had a stable draft of a survey, members of our research team translated surveys manually and compared those versions to translations via paid human translation services for robustness. We pilot tested translations with 2-4 people for each language and revised the survey further. Our goal was to have similar wording across languages; though this resulted in some overlapping terms in the prompts (e.g. malicious), participants seemed to comprehend each prompt in our pilots. We deployed the survey in a dominant local language for each country (see Table 1). The survey contained four parts: harassment scenarios, harm measures, possible remedies, and demographics and values. Below, we describe each stage in detail:

Harassment scenarios. We selected four online harassment scenarios to capture participants' perceptions about a range of harassment experiences but without making the survey too long which leads to participant fatigue. We selected the four harassment scenarios by reviewing prior scholarly literature, reports, and news articles and prioritizing diversity in types of harm and severity of harm. We prioritized harassment types that would be globally relevant and legible among participants and could be described succinctly. Participants were presented with one scenario along with the harm and remedy questions (described below), and completed this sequence four times for each harassment scenario. The harassment scenario prompt asked participants to "Imagine a person has:" and then presented each of the experiences below.

- spread malicious rumors about you on social media
- taken sexual photos of you without your permission and shared them on social media
- insulted or disrespected you on social media
- created fake accounts and sent you malicious comments through direct messages on social media

Harm measures. We developed four measures of harm to ask about with each harassment scenario. We again prioritized types of harmful experiences that would be relevant to participants globally. Drawing on our literature review on harms in other disciplines (e.g. medicine) and more nascent discussions of technological harms (e.g. privacy harms [20]), we chose to prioritize three prominent categories of harm used in scholarly literature and by the World Health Organization–psychological, physical, and sexual harm. We then added a fourth category—reputational harm—because harm to family reputation is a prominent concern in many cultures and these concerns may be exacerbated on social media. We prioritized question wording that could be translated and understood across languages. For example, our testing revealed that the concept of "physical harm" was confusing to participants when translated so we iterated on wording until we landed on personal safety. The final wording we used was:

- Would you be concerned for your psychological wellbeing?
- Would you be concerned for your personal safety?
- Would you be concerned for your family reputation?
- Would you consider this sexual harassment against you?

Perceived harm options were presented on 5-point scales of "Not at all concerned" (1) to "Extremely concerned" (5) for the first three

questions and "Definitely not" (1) to "Definitely" (5) for the last question. We chose these response stems to avoid Agree/Disagree options which may promote acquiescence bias [86] and because these could be translated consistently across languages.

Harassment remedies. Current harassment remedies prioritize content removal and user bans after a policy violation. However, scholars are increasingly arguing that a wider range of remedies is needed for addressing widespread harms. Goldman proposes that expanded remedies can improve the efficacy of content moderation, promote free expression, promote competition among Internet services, and improve Internet services' community-building functions [39]. Goldman's taxonomy of remedies is categorized by content regulation, account regulation, visibility reductions, monetary, and "other." Schoenebeck et al [88] have also proposed that expanding remedies can create more appropriate and contextualized justice systems online. They see content removal and user bans as a form of criminal legal moderation, where harmful behavior is removed from the community, and propose adding complementary justice frameworks. For example, restorative justice suggest alternative remedies like apologies, education, or mediation. Building on this work, we developed a set of proposed remedies and for each harassment scenario, we asked participants, "How desirable would you find the following responses?" with response options on a 5-point scale of "Not at all desirable for me (1)" to "Extremely desirable for me (5)." The seven remedies we displayed were chosen to reflect a diversity of types of remedies while keeping the total number relatively low to reduce participant fatigue. We also asked one free response question "What do you think should be done to address the problem of harassment on social media?"

- removing the content from the site.
- labeling the content as a violation of the site's rules.
- banning the person from the site.
- paying you money.
- requiring a public apology from the person.
- revealing the person's real name and photograph publicly on the site.
- by giving a negative rating to the person.

Demographics. The final section contained social media use, values, and demographic questions. The values and demographic questions were derived from the World Values Survey (WVS) [51], a long-standing cross-country survey of values. This paper focuses on six measures from the WVS.

- Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?
- How much confidence do you have in police?
- How much confidence do you have in the courts?
- How secure do you feel these days in your neighborhood?
- What is your gender?
- Have you had any children?

The response options ranged from "None at all" (1) to "A great deal" (4) for police and courts and from Not at all secure (1) to Very secure (4) for neighborhood. We omitted the police and courts questions in Saudi Arabia. For trust, options were "Most people can be trusted" (1) and "Need to be very careful" (2). For gender, the response options were "Male", "Female", "Prefer not to disclose",

Table 1: Participant demographics

Country	Language	Num Participants
Austria	German	251
Cameroon	English	263
Caribbean	English	254
China	Mandarin	283
Colombia	Spanish (Colombian)	296
India	Hindi/English	277
South Korea	Korean	252
Malaysia	Malay	298
Mexico	Spanish (Mexican)	306
Mongolia	Mongolian	367
Pakistan	Urdu	302
Russia	Russian	282
Saudi Arabia	Arabic	258
USA	English	304
Total		**3993**

and "Prefer to self-describe." We chose not to include non-binary or transgender questions because participants in some countries cannot safely answer those questions, though participants could choose to write them in.

We recruited participants from 14 countries (see Table 1): 13 countries plus the Caribbean countries (Antigua and Barbuda, Barbados, Dominica, Grenada, Jamaica, Monserrat, St. Kitts and Nevis, St. Lucia, and St. Vincent). We decided to analyze the Caribbean countries together because of the small sample sizes and their relative similarities, while recognizing that each country has its own economics, culture and politics. This study was exempted from review by our institution's Institutional Review Board. Participants completed a consent form in the language of the survey. Participants were recruited via the survey company Cint in most countries, Prolific in the U.S., and manually via the research team in the Caribbean countries and Mongolia. Participants were compensated based on exchange rates and pilot tests of time taken in each country.

4.1 Participant Demographics

The gender ratio between men and women participants was similar across countries ranging from 50% women and 50% men in China to 43% women and 57% men in India) except for Caribbean countries which was women: 69%, men: 27% and Mongolia which was women: 59%, men: 41% (see details about gender in [50]). The median age was typically in the 30s; Mongolia was lowest at 21 while South Korea and United States were 41.5 and 44, respectively. Participants skew young but roughly reflect each country's population, e.g. Mongolia's median age is 28.2 years while South Korea and U.S medians are 43.7 and 38.3, respectively, according to United Nations estimates [71]. Participants' self-reported income also varied across countries, with participants in Austria reporting higher incomes and participants in Caribbean countries reporting lower incomes. More than half of the participants had education equivalent to a Bachelor degree for eight countries (Cameroon, China, Colombia, India, Malaysia, Russia, Saudi Arabia, United States); the other

countries did not. Participants placed their political views as more "left" than "right."

4.2 Data analysis

We discarded low-quality responses based on duration (completed too quickly) and data quality (too many skipped questions). Table 1 shows the final number of participants per country after data cleaning. For the qualitative analysis, we separately discarded responses that were low quality (empty fields, meaningless text); the number of participants was slightly higher overall (N=4127) since some participants completed that section but did not finish the subsequent quantitative portions of the survey.

We analyzed data using R software. We used group means to describe perceived harms and preferred remedies. Levene's tests to measure variance were significant for both harm and remedy analyses indicating that homogeneity of variance assumption is violated. Thus, we used Welch one-way tests for nonparametric data and posthoc pairwise t-tests which we deemed appropriate given our sufficiently large sample size [33]. We used the Benjamini–Hochberg (BH) test to correct for multiple comparisons [12]. We also ran linear regressions with harassment - harm and harassment - remedy pairings as the dependent variables and demographics and country as the independent variables (4 harassment scenarios x 4 harm types = 16 harm models; 4 harassment scenarios x 7 remedies = 28 remedy models). We used adjusted R-squared to identify demographic variables that were more likely to explain model variance. Welch test and posthoc tests for harm (16 harassment-harm pairings) and remedy (28 harassment - remedy pairings) comparisons are available in the Appendix. Regression outputs and confidence intervals for demographic predictors are also available in the Appendix.

We analyzed the qualitative data to the free responses question using an iterative, inductive process. Our approach was to familiarize ourselves with the data, develop a codebook, iteratively refine the codebook, code the data, then revisit the data to make sense of themes from the coding process. To do this, four members of the research team first read through a sample of responses across countries and then co-developed a draft codebook. Three members of the team then coded a sample of responses and calculated interrater reliability (IRR) for each code using Cohen's Kappa. Across the 26 codes tested, Kappa values ranged from -0.1 to 1 with a median of .35. We used the IRR values as well as our manual review of differences to refine the codebook. We removed codes that coders did not interpret consistently, generally those with agreement below about .4 and those that were low prevalence in the data. We revised remaining codes, especially those that had lower agreement, and discussed them again. The final codebook contained 21 codes (see Appendix) that focused on moderation practices, user responsibility, government involvement, and other areas of interest.

4.3 Limitations and Mitigation

Cross-country surveys are known to have a range of challenges that are difficult to overcome completely, but they remain useful, even indispensable, if designed and interpreted thoughtfully and cautiously [57, 58, 92].

In our case, the key issues have to do with language, sampling methodologies, and response biases that might have differed across our participants. Language differences were addressed as described above, through a process of careful translation, validation through back-translation, and survey piloting, but topics like non-consensual image sharing are inevitably shaped by the language they are discussed in and there may be differences in interpretation we did not capture. Sampling methodologies within countries were as consistent as we could make them, but a number of known differences should be mentioned: First, we used three different mechanisms for recruiting – a market research firm (Cint) for 11 countries; a research survey firm (Prolific) for the United States; and our own outreach for the Caribbean and Mongolia. These mechanisms differ in the size of their pool of participants, as well as their baseline ability to draw a representative sample. Some differences were built-in to the recruitment process, for example, we requested a diverse age range of participants explicitly with Cint and Prolific which should have yielded more older adults. In contrast, our researcher recruitment method for Caribbean and Mongolia simply sought a range of participants through word of mouth, but did not specifically recruit or screen for older adults. Second, while we sought representative samples of the national/regional population in all cases, we know that we came up short. For example, while online surveys are increasingly able to achieve good representation in better-educated countries with high internet penetration, they are known to be skewed toward affluent groups in lower-income, less-connected contexts [69, 96]. Oversampling from groups who are active online may be more tolerable for a study of online harassment, but it still overlooks important experiences from those who may be online but less likely to participate in a survey. Third, differences in local culture and current events are known to cause a range of response biases across countries. Subjective questions about perception of harm, for example, might depend on a country's average stoicism; questions about "trust in courts" might be affected by the temporary effects of high-profile scandals. The issues above are common to cross-country survey research, and our mitigation strategies are consistent with the survey methodology literature [57, 58].

To provide some assurance of our data's validity, we benchmarked against the World Values Survey, on which some of our demographic and social-issues questions were based. We compared responses from our participants to responses from the WVS for countries that had WVS data (China, Colombia, partial India, South Korea, Malaysia, Mexico, Pakistan, Russia, United States). We used the more recent Wave 7 (2017-20) where data was available, with Wave 6 (2010-14) as a back-up. We expected that our responses should correlate somewhat with WVS, even though there were substantial differences, such as that our sample was recruited via online panels with questions optimized for mobile devices whereas the WVS sample was recruited door-to-door with oral question and answer choices. Sample means for our data and the WVS for similar questions are presented in plots in the Appendix. In countries where corresponding data is available, we find that the means in our data about trust – in police, or in courts – align with WVS results. We also find the anticipated biases with respect to online surveys and socio-economic status. In particular, our participants reported better health and more appreciation for gender equality than WVS participants.

Still, because of the above, we present our results with some caution, especially for between-country comparisons; specific pairwise comparisons between countries should be considered with substantial caution. We include specific comparisons primarily in the Appendix for transparency; we focus on patterns in the Results which we expect to be more reliable, especially patterns within countries and holistic trends across the entire dataset. In the following sections, we strive to be explicit about how our findings can be interpreted.

5 RESULTS

Results are organized into two sections: perceptions of harm associated with online harassment and preferences for remedies associated with online harassment. Each section follows the same structure: first we look at which harassment types are perceived as most harmful and which remedy types are most preferred, respectively, then we examine demographic predictors of perceptions of harm and preferences for remedies, respectively.

5.1 Perceptions of Harm Associated with Online Harassment

First, we differentiate between the four types of harassment. Figure 1 shows perceptions of overall harm by harassment type. One-way Welch tests showed that means of perceptions of harm were significantly different, $F(3, 35313) = 3186.4$, $p < 0.001$, with sexual photos being the highest in harm (M=4.20, SD=1.15), followed by spreading rumors (M=3.42; SD=1.30), malicious messages (M=3.20; SD=1.35), and insults or disrespect (M=2.93; SD=1.36) (see Figure 1). Plots and posthoc tests for comparisons by type of harassment by country are available in the Appendix.

To display an overall measure of perceived harms associated with online harassment by country, we aggregated each of the four harm measures together – sexual harassment, psychological harm, physical safety, and family reputation – for a combined measure of overall harm.

Results suggest that participants in Colombia, India, and Malaysia rated perceived harm highest, on average, while participants in the United States, Russia, and Austria perceived it the lowest. Means are presented here and shown visually in Figure 2: Colombia (M=3.98; SD=1.18); India (M=3.86; SD=1.31); Malaysia (M=3.79; SD=1.21); Korea (M=3.67; SD=1.22); China (M=3.59; SD=1.19); Mongolia (M=3.55; SD=1.29); Cameroon (M=3.50; SD=1.38); Caribbean (M=3.44; SD=1.43); Mexico (M=3.38; SD=1.38); Pakistan (M=3.36; SD=1.36); Saudi Arabia (M=3.34; SD=1.35); Austria (M=2.99; SD=1.42); Russia (M=2.80; SD=1.43); United States (M=2.79; 1.45).

Most of the ratings by country were statistically significant from each other (one-way Welch tests and posthoc tests are reported in the Appendix), though we remind readers that these differences should be interpreted with caution. In general, the wealthier countries per capita perceive lower harm, but beyond that the key takeaway is that there is substantial variance which is unlikely to be explained by one or even a few differences across any of those countries.

5.1.1 Predictors of Perceptions of Harm. Here we hone in more granular differences across harassment types and harm types and how they vary by country and other demographic data. Note that
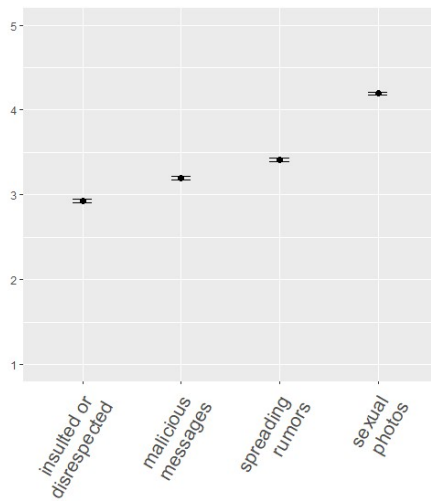
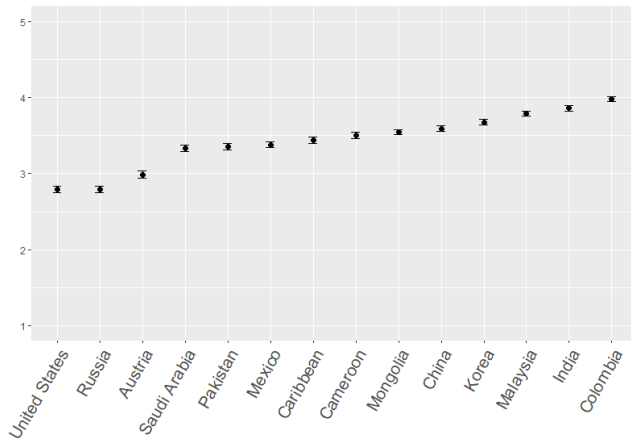**Figure 1: Perceptions of harm by harassment type**

Figure 2: Perceptions of harm by country

responses from Saudi Arabia participants are excluded from regressions because they did not complete questions about confidence in courts or police. The distribution of R-squared values for the 16 harassment - harm pairings is shown in Figure 3 (ranging from close to 0 to 18% variance). Country was the most predictive of perception of harm though with variance across harassment and harm pairings as indicated by the multiple peaks in Figure 2. Gender was next most predictive, followed by security in neighborhood, number of children, trust of people, trust in courts, and trust in police.

We also ran exploratory factor analyses to look for underlying constructs across measured variables. When all variables we measured were in the analysis, perceptions of harm and preferred remedies loaded into constructs, as expected, but demographic and value variables did not. Analyses with only the demographic and values variables suggest some trends but they were not substantial predictors of variance (e.g. trust and courts loaded together;

marriage and age inversely loaded together). We show some factor analyses results in the Appendix but do not focus on them further.
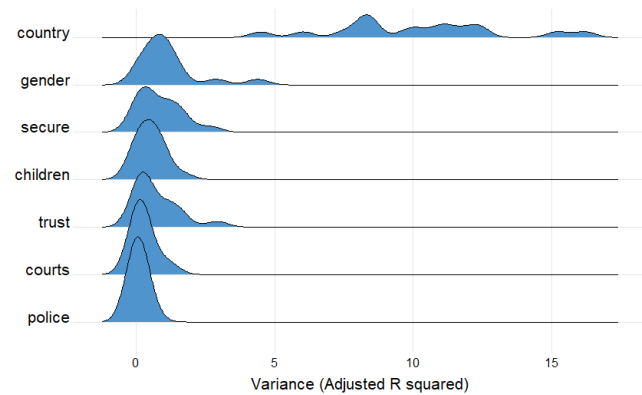

Figure 3: Adjusted R squared of demographic variables for predicting harm across 16 harassment scenario x harm types.

We ran regression analyses for the 16 harassment type - harm pairings using country, gender, security in neighborhood, number of children, trust in other people, trust in courts, and trust in police as independent variables. We used the U.S. as the reference choice for country and men as the reference for gender. Complete results with confidence intervals are available in the Appendix. To communicate patterns across models, we present a heatmap (see Figure 4) of regression coefficients with harassment type - harm pairings on the x-axis and the predictors in Figure 3 on the y-axis. We also plotted participant responses to the courts, police, security, and trust questions with WVS ratings to benchmark that our participants' attitudes reflect those of a broader population; those plots are in the Appendix.

Results of the six predictors in the regression models are summarized here:

Country: Participants in most countries perceive higher harm for most pairings than the U.S., with the exception of the sexual photos and sexual harm pairing where some countries perceive lower harm than the U.S.

Gender: Women perceive greater harm than men for all 16 harassment - harm pairings.

Secure: Participants who were more likely to give low ratings to the question "How secure do you feel these days in your neighborhood?" were more likely to perceive higher harm associated with online harassment for 8 of the harassment - harm pairings; however, security in neighborhood is negatively correlated with ratings for the sexual photos - sexual harassment pairing.

Children: Having more children is a predictor of greater perceptions of harm for 9 of the 16 pairings, except for the insulted or disrespected - sexual harassment pairing which is negatively correlated.

Trust: Participants who were more likely to be low in trust of other people were more likely to perceive higher harm associated with online harassment for 11 of the 16 harassment - harm pairings.
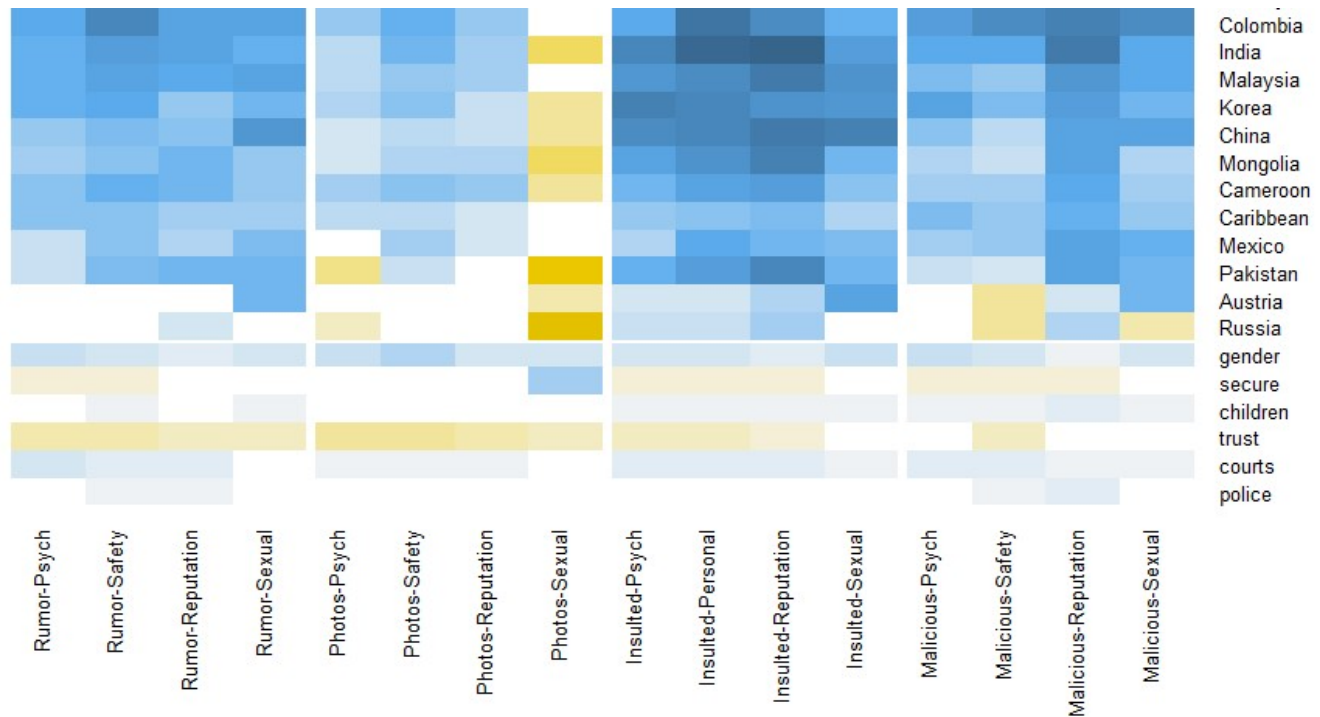
Figure 4: Heatmap of regression coefficients of harassment types and harm pairings by country and demographics. Darker blue is positive coefficient (i.e. higher harm); darker gold is negative coefficient.

The relationship was stronger for the sexual photos and spreading rumors scenarios, whereas there were no relationships for the malicious harassment scenario.

Courts: Higher trust in courts is correlated with increases in perceptions of harm for 14 of the 16 pairings. The two exceptions are spreading rumors - sexual harassment and sexual photos - sexual harassment pairings.

Police: Trust in police is correlated with increases in perceptions of harm for 4 of the 16 pairings.

We return to these results in the Discussion.

5.2 Preferences for Remedies Associated with Online Harassment

The prior section presented perceptions of harm; this section presents preferences for remedies. Specifically we report respondents' perceived desirability of the remedies to address harassment - related harms.

First, we differentiate between the remedies themselves. One-way Welch tests showed that means of preferences for remedies were significantly different, $F_{(6, 49593)} = 1130.9$, $p < 2.2e-16$ (see Figure 5). Removing content and banning offenders are rated highest, followed by labeling, then apologies and rating. Revealing identities and payment are rated lowest. Posthoc comparisons showed that all pairings were significantly different from each other except for apology and rating: removing (M=4.18; SD=1.12); banning (M=4.07; SD=1.17); labeling (M=4.00; SD=1.19); apology (M=3.72;

SD=1.34); rating (M=3.72; SD=1.32); revealing (M=3.56; SD=1.39); paying (M=3.16; SD=1.46).

To display overall preferences for remedies associated with online harassment by country, we aggregate the seven remedy types together for a combined measure of overall remedies. Results show that Colombia, Russia, and Saudi Arabia were highest overall in support for remedies while Pakistan, Mongolia, and Cameroon were lowest. Means are again presented here and shown visually in Figure 6: Colombia (M=4.07; SD=1.13); Russia (M=4.03; SD= 1.25); Saudi Arabia (M=3.97; SD=1.27); Mexico (M=3.93; SD=1.25); Malaysia (M=3.89; SD=1.20); China (M=3.89; SD=1.07); Caribbean (M=3.86; SD=1.38); Austria (M=3.86; SD=1.36); Korea (M=3.72; SD=1.29); India (M=3.70; SD=1.39); United States (M=3.60; SD=1.50); Cameroon (M=3.57; SD=1.40); Mongolia (M=3.45; SD=1.40); Pakistan (M=3.40; SD=1.41). As with harms, most of the ratings by country were statistically significant from each other (one-way Welch tests and posthoc tests are reported in the Appendix), though we again caution that the differences should be treated with caution. Regression outputs and confidence intervals for demographic predictors are also available in the Appendix.

5.2.1 Predictors of Preferences for Remedies. We plotted R-squared values with the same variables used in the harm regression models (see Figure 7). Results are broadly similar to the harm ridgeline plot, though there is less overall variance explained in the remedy plots (0-15%). Country is most predictive of preference for remedy, followed by number of children, gender, security in neighborhood, trust in police, trust in courts, and trust in other people. We ran
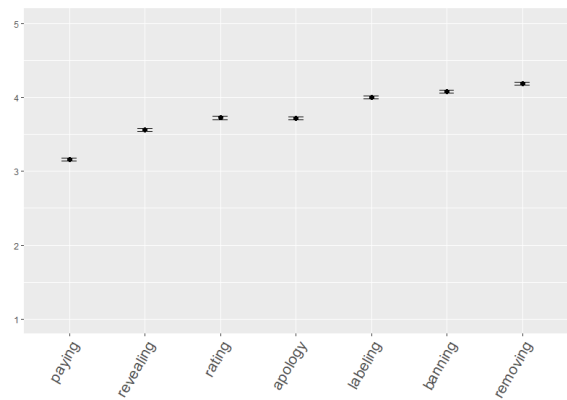
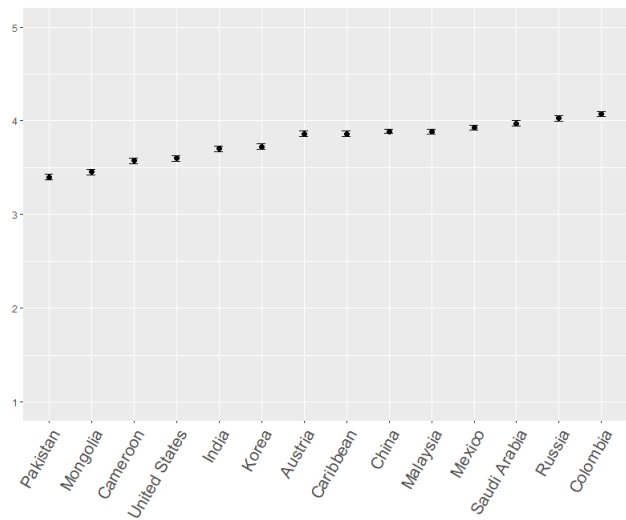**Figure 5: Preferences for remedies by remedy type**

Figure 6: Preferences for remedies by country

regression analyses for the 28 harassment type - remedy pairings (4 harassment types and 7 remedies).

We visually show model results in a heatmap (see Figure 8). Results are summarized here:

Country: Most countries tend to prefer payment, apologies, revealing users, and rating users, but are less favorable towards removing content, labeling content, or banning users compared to the U.S. These patterns are observed for three of the four harassment types, with the exception of insults or disrespect where countries tend to prefer all remedies compared to the U.S.

Gender: Women tend to prefer most remedies compared to men, except for payment, which they are less favorable towards for all four harassment types.

Children: Having more children is associated with higher preferences for most remedies.

Secure: Security in neighborhood is negatively associated with higher preference for remedies for 8 of the 28 pairings, primarily for removing content and labeling content.
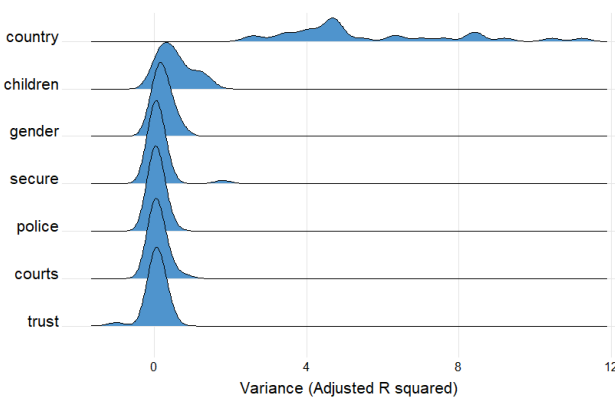

Figure 7: Adjusted R squared of demographic variables for remedy preferences across 28 harassment scenario x remedy types.

Trust: Trust in other people is negatively associated with preferences for remedies for 19 of the 28 pairings.

Courts: Confidence in courts is associated with preference for the payment remedy for all harassment types but few other remedies.

Police: Confidence in police is not correlated with remedy preferences.

5.3 Qualitative responses

We asked participants one free response question about how harassment on social media should be addressed. The most prevalent code related to the site being responsible for addressing the problem; nearly 50% of responses referred to site responsibility, ranging from about 20% to 60% across countries. This was most prevalent in Malaysia, Cameroon, and the United States and least prevalent in Mongolia. Responses included content about setting policies, enforcing policy, supporting users, and protecting users. For example, a participant in Korea said: "The social media company is primarily responsible for it, though the person who harassed others also has responsibility quite a lot." As part of this site responsibility, many participants described what they thought the site should do, such as: "The social media site should give the offender a negative rating and ban them for a specific time period. This time period being months or years or indefinitely; as well as disallowing them from creating further accounts" (Caribbean). Some participants described why they thought social media sites were responsible, such as this one from Pakistan who said: "This problem should be solved only by the social media website as they have all the data of the user through which they can take action against it."

The second most prevalent code, found in nearly 25% of responses, referred to government involvement, including regulation, police, courts, arrest, prison, criminal behavior, and juries. References to government involvement were highest in China and Pakistan and lowest in Russia, Mexico, and Cameroon. Many participants who mentioned government responsibility for online harassment indicated it should be in collaboration with law enforcement. For example, one participant from Malaysia said: "The responsible party (social media workers) need to take this issue seriously and
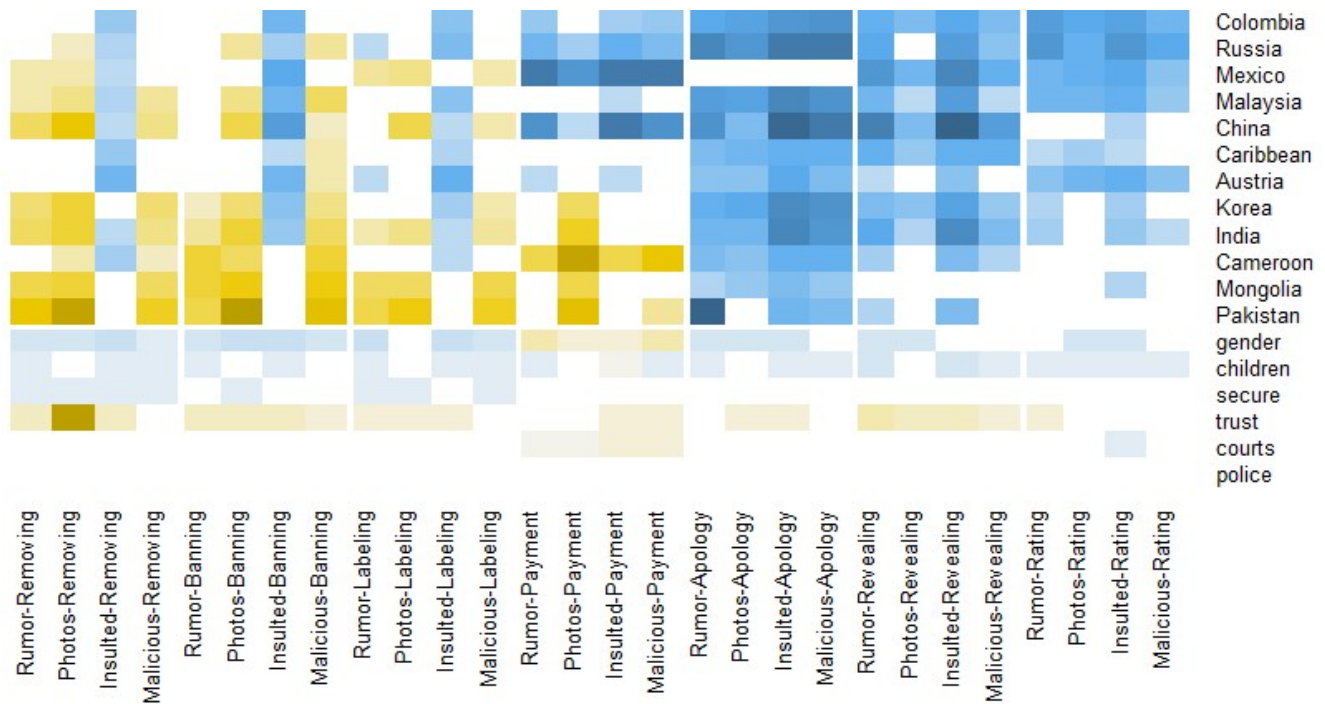
Figure 8: Heatmap of regression coefficients of harassment types and remedy pairings by country and demographics. Darker blue is positive coefficient (i.e. higher preference for remedy); darker gold is negative coefficient.

take swift action to 'ban' the perpetrator and bring this matter to court and the police so that the perpetrator gets a fair return. This is to avoid the trauma suffered by the victim and also reduce mental illness." Participants varied in their indication of whether the user or the platform should report the behavior to the government. One in India said "First of all the person who has been harassed he should simply go to police station to report this incident then he should report the account and telling his friends to report it then he should mail to instagram some screenshots of that person's chat." Some participants focused only on government responsibility, such as one in Colombia saying: "Having permission for the police to see all your material on the network and electronic devices."

References to managing content (e.g. removing or filtering content) and account sanctions (e.g. warnings, suspensions, bans) showed up in about 20% of responses. These were highest in the Caribbean and the United States and lowest in China and Mongolia. Sometimes posts only recommended one step, like content removal, but more often they mentioned a multi-step process for accountability. A participant in China recommended a user rating approach: "in serious cases, he should be banned and blacklisted, and the stains of his behavior should be recorded on his personal file." Many participants proposed policies that required real identities linked to the account to deter harassment and allow punishment. One person in Austria said: "Login to the networks only with the name known to the operator. Strict consequences, first mark the content, then also delete the profile and transfer the data to the public prosecutor's office."

About 13% of responses referred to user responsibility, in which the person who may be experiencing or targeted by harassment should handle it themselves. These responses suggested that people should ignore harassment, stop complaining or whining about it, deal with it, and understand people will disagree. These responses were highest in Colombia and Malaysia and lowest in Austria, Cameroon, Caribbean. For example a participant in Colombia indicated that users should take steps to protect themselves: "1. Being on social media is everyone's responsibility. 2. In social networks, you should limit yourself in admitting requests from strangers. 3. Remove and block malicious invitations." One in Malaysia indicated that people should work out the problems themelves: "All parties must cast off feelings of hatred and envy towards others. To deal with this problem all parties need to be kind to each other and help each other." Another Malaysia participant was more explicit, saying: "The responsible party is myself, if it happens to me, I will definitely block the profile of the person who is harassing me. The self is also responsible for not harassing others despite personal problems. It is best to complete it face to face."

Some responses, about 7%, referred to public awareness or public shaming as a response, which could be through formal media coverage or offender lists or through informal user behaviors. These were highest in the Caribbean and China. One participant in Mongolia said: "This role belongs to the private organization that runs the first social platform and to the police of the country. Disclosure to the public of the crimes committed by the perpetrators and the number of convictions related to this issue, and the damage caused by the crime."

About 6% of responses directly addressed the offender's role in the harassment, indicating that they are responsible and should address the problem and change their behavior. This was most prevalent in Korea and Malaysia. About 6% referred to restitution in some way, which could involve the offender paying a fine to the site or the victim or the victim receiving compensation from the site or offender. About 6% referred to blocking other users as a remedy.

Other codes showed up in around 5% of responses, including verifying accounts (checking for bots, fake accounts), educating users about appropriate behaviors, and offender apologies for behavior. Apologies, when specified, were often supposed to be public rather than private, such as a participant from Cameroon's response: "They should demand a public apology from the person, and if the person does not give the apology, the account should be banned." A person in China associated the public apology with reputational damage: "Infringement of my right of reputation requires a public apology to compensate for the loss." In terms of account verification, some participants talked about real names or use of IPs. One from Colombia said: "Social networks should block people's IPs and allow them to have a maximum of 2 accounts per IP, since most of the people who harass do not do it from their main accounts but rather hide inside other personalities that are not them, by Doing this would greatly reduce this type of bullying."

6 DISCUSSION

Our findings coalesce into three broad themes about global perceptions of social media harassment harms and remedies: (1) Location has a large influence on perceptions. (2) The causes are complex – no single factor, nor even a straightforward subset of factors, emerges as a dominant predictor of perceptions of harm. (3) One-size-fits-all approaches to governance will overlook substantial differences in experiences across countries.

6.1 Key Role of Local Cultural Context

Our results suggest that local cultural context plays the greatest role in determining people's perceptions of online harassment among the factors we measured. In our analysis, country emerged as the most predictive of perceptions of harm across harm types and also with respect to remedies. This is striking, especially when considering that country explained more of the variation in perceptions than gender. As is widely understood, women and girls bear a greatly disproportional brunt of harassment in general [8, 9, 14, 28, 31], and though women in each country consistently perceived greater harm than the men in the same country, women's perception of harm globally depended even more on their country. Thus for example, our data indicates that women in the United States perceive less harm from social media harassment (M=2.99, SD=1.44) than men in China (M=3.47, SD=1.2) or India (M=3.18, SD=1.33). Those are just three data points, and we do not claim that this particular set of comparisons is necessarily reliable, but it illustrates a broader point that we believe is robust from our data: *some* countries' women, on average, perceive less harm under some social media harassment cases than *other* countries' men, on average.

It is unclear what it is about local cultures that has this impact (our findings suggest that there is unlikely to be a simple set of causes), and we also wish to avoid an unresolvable discussion about what exactly constitutes "culture." Yet, it seems safe to conclude that a range of complex social factors that have some coherence at the national/regional level has a profound effect on how citizens of those countries and countries perceive social media harms and remedies. These are also inevitably shaped by policies and regulations in those countries. For example, some of our Malaysia participants said that online harassment should be the responsibility of the "MCMC." The Malaysian Communications and Multimedia Commission is responsible for monitoring online speech, including social media, though it has little power to remove content on platforms hosted outside of Malaysia. Though all countries we studied have some laws governing the extent of critique users can express towards their own governments, these laws vary in severity. For example, in 2015, the Malaysian government asked Facebook and YouTube to take down posts by blogger Alvin Tan which insulted Muslims [1]. More recently, the Indian government not only sanctioned individual users who critiqued Modi, but it sought to sanction Twitter for not taking down those posts – Twitter has recently launched a lawsuit against the Indian government in response [3]. Though insults is lower in harm than other types of harassment, it is higher in some countries in our study, and it is the most prevalent type of harassment among participants in Google's 22-country survey, suggesting that it may have cumulative harmful effects for users [95].

At the same time, it is important to remember that experiences and concerns within countries inevitably vary and span across boundaries. Our data indicates that reputational harm is lower in the U.S. and Austria and this may be true for the majority in our sample from those countries, but reputational harm can persist within and across boundaries. For instance, Arab women living in the U.S. may deal both with Arab and Western patriarchal structures and orientalism, thereby experiencing a form of intersectional discrimination that requires specific support measures and remedies [6]. Similarly, refugees and undocumented migrants may be less likely to report online harassment for fear of repercussion to their status in the country [41]. Though a focus on country-level governance is important, additional work is required to protect and support people within countries who may experience marginalization, despite or because of, local governance.

6.2 No Simple Causal Factors for Harm Perception

The second broad conclusion of our study is that perceptions of harm about online harassment are complex; no simple mechanism, nor any small set of variables, easily explains relative perceptions among countries. Harm perceptions might, for example, reasonably be expected to correlate with how much people trust others, how safe they feel in their own neighborhoods, or how much they trust institutions like the police and the courts. Yet, our results find no such easy explanations: sense of neighborhood security correlated positively with greater perceptions of harm for some forms of harassment, but negatively for the nonconsensual sharing of sexual photos and sexual harassment pairing; number of children predicted greater harm for half of the harassment - harm pairings, but not the other half.

Some correlations did emerge in our data, but it is not straightforward to interpret them. For example, trust in courts was associated with perceptions of harm in a majority of our countries. This pattern is surprising, and could indicate a desire to normalize online harassment as harmful to enable greater judicial oversight over those harms. Interestingly, trust in courts is mostly not correlated with the remedies we measured, *except* for payment which is negatively correlated. It may be that lower trust in courts to procure compensation may be correlated with a higher reliance on platforms, but we would need additional data to confirm this interpretation.

Somewhat easier to explain is that trust in other people was correlated with lower perception of harm in most cases. It may be that people who are low in trust in others assume online harassment will be severe and persistent. There was substantial variance in trust levels between countries, with Caribbean being lowest and China being highest. This suggests that harms associated with online harassment may reflect offline relationships and communities.

Our results show that there is little or no relationship between confidence in police and harm or remedies, which may indicate that people do not see online harassment as a problem that police can or should address. This interpretation aligns also with previous research which has highlighted how police are often an inadequate organization to deal with concerns around harassment and online safety, and can sometimes cause more harm [85]. Instead, experts have called for investments in human rights and civil society groups who are specifically trained to support people in the communities who experience harassment [110]. Such experts could also mediate between affected people and other institutions such as the police and legal institutions.

An exploration of factors we did not consider may find simpler or more coherent causal explanations for perceptions of harm and remedies, but we conjecture that the complexity is systemic. Online harassment, though relatively easy to discuss as a single type of phenomenon, touches on many social, cultural, political, and institutional factors, and the interplay among them is correspondingly complex. A highly patriarchal honor culture that leads women to fear the least sensitive of public exposures might be partially countered by effective law enforcement that prioritizes those women's rights; deep concerns about one's children might be offset by a high level of societal trust; close-knit communities might on the one hand provide victims with healthy support, but they might also judge and impose harsh social sanctions.

6.3 One Size Does Not Fit All in Online Governance

The four types of harassment we studied all differed from each other in perceived harm, both in type of harm and severity of that harm. Non-consensual sharing of sexual photos was highest in harm, consistent with work on sexual harms that has focused on non-consensual sharing of sexual images [20, 38, 98]. This work has advocated for legal protection and recourse for people who are victims of non-consensual image sharing and has brought attention to the devastating consequences it can have on victims' lives. Much of this transformative work in U.S. contexts focuses on sexual content like nude photos, which are now prohibited in some states in the U.S. (though there is no federal law) [19]. However, in many

parts of the world there are consequences for sharing photos of women even if they do not contain nude content.

Our findings show substantial variance in perceptions of reputational harm as well as physical harm between countries. India (medians of 4.09 and 4.01, respectively) and Colombia (4.02, 4.24) are highest in both of those categories whereas the U.S. is lowest (2.73, 2.69). Our results corroborate Microsoft's Digital Civility Index, which found high rates of incivility in Colombia, India, and Mexico (and the U.S. being relatively low), though Russia was also high which deviates from our results. Google's survey similarly shows Colombia, India, and also Mexico as highest in prevalence of hate, abuse, and harassment [95]. While shame associated with reputation persists globally, it may be a particularly salient factor where cultures of honor are high [82]. In qualitative studies conducted in the South Asian country, including India, Pakistan, and Bangladesh, participants linked reputational harm with personal content leakage and impersonation, including non-consensual creation and sharing of sexually explicit photos [85]. Because women in conservative countries like India are expected to represent part of what the family considers its "honor," reputational harm impacts not only just the individual's personal reputation but also their family and community's reputation.

As one South Asian activist described technology-facilitated sexual violence (quoted from [62]): *"A lot of times, there's an overemphasis on sexually explicit photos. But in [this country], just the fact that somebody is photographed with another boy can lead to many problems, and we've seen honor killings emerging from that."* In these cases, women are expected to represent part of what the family considers its "honor" [85] and protecting this honor becomes the role of the family, and especially men in the family, who seek to regulate behavior to preserve that honor. Unfortunately, when a person becomes a victim of online abuse, it becomes irrelevant whether she is guilty or not, what matters is other people's perception of her guilt. At an extreme, families will engage in honor killings of women to preserve the honor of the family [46, 56, 80].

When women experience any kind of abuse, they may need to bring men with them to file reports, and then they may be mocked by officials who further shame and punish them for the abuse they experienced [85]. In Malaysia, legal scholars raise concern about the inadequacy of law in addressing cyberstalking in both the National Cyber Security Policy and the Sustainable Development Goals [83]. Sexual harassment, sexual harm, and reputation are strongly linked, and the threat of reputational damage empowers abusers. Many European countries have taken proactive stances against online harassment but the efficacy of their policies are not known yet. Unfortunately, any efforts to regulate content also risk threats to free-expression, such as TikTok and WeChat's suppression of LGBTQ+ topics [101]. Concerns about human rights and civil rights may be especially pronounced in countries where there is not sufficient mass media interest to protest them, such as the rape of a girl in India by a high-profile politician that did not gather attention because it was outside of major cities [42].

In Latin American contexts, there is similar evidence that societies that place a premium on family reputation are likely to be afflicted by higher rates of intrapersonal harm [25, 72]. For example, constitutional laws against domestic violence in Colombia decree that family relations are based on the equality of rights and

duties for all members, and that violations are subject to imprisonment. Yet recent amendments have called for retribution against domestic violence to be levied *only* when charges with more severe punishment do not apply. Human rights activists from the World Organisation Against Torture have claimed that such negligent regulations send the message that domestic violence, including harassment, is not as serious as other types [79, 97]. Even with the existence of laws on domestic violence in countries like Colombia and Mexico, prevailing attitudes view harassment as a "private" matter, perhaps because of traditional norms that value family cohesion over personal autonomy. One speculation is that fears of reporting harassment because of familial backlash corroborate why survey respondents from this country may not find exposing their abusers online satisfying.

6.4 Recommendations for Global Platform Design and Regulation

Our recommendations for global platform design and regulation build on work done by myriad civil society groups and follow from our own findings. In short, harms associated with online harassment is greater in non-U.S. countries and platform governance should be more actively coshaped by community leaders in those countries. Above all, we discourage any idea that a *single* set of platform standards, features, and regulations can apply across the entire world. While a default set of standards might be necessary, the ideal would be for platforms and regulations to be further customized to local context. A reasonable start is for platforms to regulate at the country level, though governance should be sensitive to the blurriness of geopolitical and cultures boundaries. Digital technology is highly customizable, and it would be possible to have platform settings differ by country. Similarly, regulation of social media, as well as policy for harm caused through online interaction, should also be set locally. To a great extent the latter already happens, as applicable policy tends to be set at a national level. It should also be the case that technology companies engage with local policymakers, without assuming that one-size-fits-all approaches are sufficient.

According to the findings discussed above, local cultural context can play an important role in helping platforms define harassment and prioritize online speech and behavior that will likely have the most impact in a given local context. For example, posting nonconsensual images, whether sexual or not, can have a more severe impact in countries where women's visibility and autonomy are contentious issues. Customizing definitions of harms would also align with the task of determining the effectiveness of a remedy. If certain behaviors are criminalized offline, that would likely have an impact on how seriously platforms should take online manifestations of such harassment, and how easy it would be for users in that locality to seek help from police or courts. Lastly, due to the great variance on how local laws are shaped and implemented, platforms can play a key role in determining the effectiveness of rules as applied to them and their users. The resulting observations about what laws are effective on the ground can help platforms both customize their own policies, and engage with stakeholders more productively.

Platform features, settings, and regulation ought to be determined by multistakeholder discussions with representation from local government, local civil society, researchers, and platform creators. Input from entities familiar with the local laws, customs, and values is essential, as others have recommended (e.g. [15, 110]). As our study also finds, the specifics of how users respond to online harassment are localized and not given to easily generalized explanation. Of course, such discussions must be designed well. For example, we recommend that platform creators – who have international scope yet often tend toward Western, educated, industrial, rich, and democratic (WEIRD) sensibilities [47, 61] – take a back seat and turn to local community leaders to lead these discussions. Platform creators have the power to determine final features anyway; additional exertion of power in such discussions will suppress local voices. Tech companies must also be willing to adopt the resulting recommendations [77].

Beyond platform and regulatory customization within countries, there should be transnational bodies that consider things at a global level, and which might also serve to mediate between issues that bring geographic countries into contention. Technology companies already sponsor such bodies – for instance, Meta has a Stakeholder Engagement Team that includes policymakers, NGOs, academics, and outside experts that support the company in developing Facebook community standards and Instagram community guidelines [67]. Even better would be for such bodies to have more independence, set up for autonomous governance via external organizations.

We recognize that customization by country raises new challenges, such as the question of whose policy should take precedence when cross-country interaction occurs on a platform. Or, how platforms should handle users who travel across countries (or claim to do so). Or the substantial problem, though not the focus of this paper, of how to address authoritarian regimes that are not aligned with human rights [110]. It will take work, and diplomacy, to resolve these issues, but if the aim is to prevent or mitigate harassment's harms in a locally appropriate way, the effort cannot be avoided. As to what kinds of customization such bodies might suggest, our study gestures toward features and regulations that might differ from place to place. For example, there appears to be wide variation across countries in terms of what is considered invasive disclosure. Russians generally care much less than Pakistanis whether photographs of an unmarried/unrelated man and woman are posted publicly. Thus, in some contexts, the default setting might require the explicit consent of all tagged, commented, or (automatically) recognized parties for a photo or comment to be posted. Another possibility is to adjust the ease with which a request to take down content is granted. The possibilities span a range from (A) automatically taking down any content as requested by *anyone* to (Z) refusing to take down any content regardless of the volume or validity of requests. In between, there is a rich range of possibilities that could vary based on type of content and on country. With respect to how platforms manage content-removal requests, they might establish teams drawn from each geographic context, so that decision-makers address requests from cultures they are most familiar with (and based on standards recommended by the aforementioned local bodies).

7 CONCLUSION

We studied perceptions of harm and preferences for remedies associated with online harassment in 14 countries around the world. Results show that all countries perceive greater harm with online harassment compared to the U.S. and that non-consensual sharing of sexual photos is highest in harm, while insults and disrespect is lowest. In terms of remedies, participants prefer removing content and banning users compared to revealing identities and payment, though they are more positive than not about all remedies we studied. Country is the biggest predictor of ratings, with people in non-U.S. and lower income countries perceiving higher harm associated with online harassment in most cases. Most countries prefer payment, apologies, revealing identities, and rating users compared to the U.S., but are less favorable towards removing content, banning users, and labeling content. One exception to these trends is non-consensual sharing of sexual photos, which the U.S. rates more highly as sexual harassment than other countries. We discuss the importance of local contexts in governing online harassment, and emphasize that experiences cannot be easily disentangled or explained by a single factor.

ACKNOWLEDGMENTS

REFERENCES

[1] 2015. Troll or hero? The sex blogger who's offending Muslims. (2015). https://www.bbc.com/news/blogs-trending-32515516
[2] 2020. This is how much Americans trust Facebook, Google, Apple, and other big tech companies: 2020. (2020). https://www.theverge.com/2020/3/2/21144680/verge-tech-survey-2020-trust-privacy-security-facebook-amazon-googleapple
[3] 2022. Twitter launches legal challenge in India over orders to block content. (2022). https://www.cnn.com/2022/07/06/tech/twitter-legal-action-india-government-blocked-content-intl-hnk/index.html
[4] Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society.* 298–306.
[5] Sara Ahmed. 2021. *Complaint!* Duke University Press.
[6] Ali J Al-Kandari, Ahmed A Al-Hunaiyyan, and Rana Al-Hajri. 2016. The influence of culture on instagram use. *Journal of Advances in Information Technology,* 7 (1), 54 57 (2016).
[7] Lenhart Amanda, Ybarra Michele, Zickuhr Kathryn, and Myeshia Price-Feeney. 2016. Online harassment, digital abuse, and cyberstalking in America. *Data and Society Research Institute* (2016).
[8] Yara Barrense-Dias, Christina Akre, Diane Auderset, Brigitte Leeners, Davide Morselli, and Joan-Carles Surís. 2020. Non-consensual sexting: Characteristics and motives of youths who share received-intimate content without consent. *Sexual Health* 17, 3 (2020), 270–278.
[9] Samantha Bates. 2017. Revenge porn and mental health: A qualitative analysis of the mental health effects of revenge porn on female survivors. *Feminist Criminology* 12, 1 (2017), 22–42.
[10] Laura-Kate Bernstein. 2016. Investigating and prosecuting swatting crimes. *US Att'ys Bull.* 64 (2016), 51.
[11] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and its consequences for online harassment: Design insights from heartmob. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–19.
[12] Frank Bretz, Torsten Hothorn, and Peter Westfall. 2016. *Multiple comparisons using R.* Chapman and Hall/CRC.
[13] Katharina Buchholz. 2022. How has the number of female CEOs in Fortune 500 companies changed over the last 20 years? (2022).
[14] Melissa Burkett. 2015. Sex (t) talk: A qualitative analysis of young adults' negotiations of the pleasures and perils of sexting. *Sexuality & culture* 19, 4 (2015), 835–863.
[15] Bart Cammaerts and Robin Mansell. 2020. Digital platform policy and regulation: Toward a radical democratic turn. *International journal of communication* 14 (2020), 20.
[16] Pew Research Center. 2021. The State of Online Harassment. https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/
[17] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017. The bag of communities: Identifying abusive behavior online with preexisting internet data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems.* 3175–3187.
[18] Sarai Chisala-Tempelhoff and Monica Twesiime Kirya. 2016. Gender, law and revenge porn in Sub-Saharan Africa: a review of Malawi and Uganda. *Palgrave Communications* 2, 1 (2016), 1–9.
[19] Danielle K Citron and Mary Anne Franks. 2019. Evaluating New York's 'Revenge Porn'Law: A missed opportunity to protect sexual privacy. *Harvard Law Review Blog* 19, 3 (2019).
[20] Danielle Keats Citron and Daniel J Solove. 2021. Privacy harms. *Available at SSRN* (2021).
[21] Danielle J Corple. 2016. *Beyond the Gender Gap: Understanding Women's Participation in Wikipedia.* Ph. D. Dissertation. Purdue University.
[22] Kate Crawford and Tarleton Gillespie. 2016. What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society* 18, 3 (2016), 410–428.
[23] Dawn. 2013. Two girls, mother killed over family video. https://www.dawn.com/news/1020576/two-girlsmotherkilled-over-family-video
[24] Isabel Debre and Fares Akram. 2021. Facebook's language gaps weaken screening of hate, terrorism. (2021). https://apnews.com/article/the-facebook-papers-language-moderation-problems-392cb2d065f81980713f37384d07e61f
[25] Dorothee M Dietrich and Jessica M Schuett. 2013. Culture of honor and attitudes toward intimate partner violence in Latinos. *Sage Open* 3, 2 (2013), 2158244013489685.
[26] Jill P Dimond, Michaelanne Dye, Daphne LaRose, and Amy S Bruckman. 2013. Hollaback! The role of storytelling online in a social movement organization. In *Proceedings of the 2013 conference on Computer supported cooperative work.* 477–490.
[27] Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2, 3 (2012), 1–30.
[28] Amy Shields Dobson and Jessica Ringrose. 2016. Sext education: pedagogies of sex, gender and shame in the schoolyards of Tagged and Exposed. *Sex Education* 16, 1 (2016), 8–21.
[29] Evelyn Douek. 2020. Governing Online Speech: From'Posts-As-Trumps' to Proportionality and Probability. *Columbia Law Review* 121, 1 (2020).
[30] Maeve Duggan. 2017. Online harassment 2017. (2017). https://ncvc.dspacedirect.org/handle/20.500.11990/10
[31] Stine Eckert and Jade Metzger-Riftkin. 2020. Doxxing. *The international encyclopedia of gender, media, and communication* (2020), 1–5.
[32] Elizabeth Englander. 2015. Coerced sexting and revenge porn among teens. *Bullying, teen aggression & social media* 1, 2 (2015), 19–21.
[33] Morten W Fagerland. 2012. t-tests, non-parametric tests, and large studies—a paradox of statistical practice? *BMC medical research methodology* 12, 1 (2012), 1–7.
[34] Casey Fiesler, Cliff Lampe, and Amy S Bruckman. 2016. Reality and perception of copyright terms of service for online content creation. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing.* 1450–1461.
[35] United Nations Entity for Gender Equality and the Empowerment of Wome. 1994. Declaration on the Elimination of Violence against Women. https://www.un.org/womenwatch/daw/vaw/reports.htm#declaration
[36] Jesse Fox and Wai Yen Tang. 2017. Women's experiences with general and sexual harassment in online video games: Rumination, organizational responsiveness, withdrawal, and coping strategies. *New Media & Society* 19, 8 (2017), 1290–1307.
[37] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media.* Yale University Press.
[38] Carrie Goldberg. 2019. *Nobody's Victim: Fighting Psychos, Stalkers, Pervs, and Trolls.* Plume.
[39] Eric Goldman. 2021. Content Moderation Remedies. *Michigan Technology Law Review, Forthcoming* (2021).
[40] Mark Griffiths. 2002. Occupational health issues concerning Internet use in the workplace. *Work & Stress* 16, 4 (2002), 283–286.

[41] Tamy Guberek, Allison McDonald, Sylvia Simioni, Abraham H Mhaidli, Kentaro Toyama, and Florian Schaub. 2018. Keeping a low profile? Technology, risk and privacy among undocumented immigrants. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–15.

[42] Pallavi Guha. 2021. *Hear# metoo in India: News, Social Media, and Anti-Rape and Sexual Harassment Activism*. Rutgers University Press.

[43] Vikram Gupta, Rini Sharon, Ramit Sawhney, and Debdoot Mukherjee. 2022. ADIMA: Abuse Detection In Multilingual Audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6172–6176.

[44] Oliver L Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 2021. Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–35.

[45] Rebecca M Hayes and Molly Dragiewicz. 2018. Unsolicited dick pics: Erotica, exhibitionism or entitlement?. In *Women's Studies International Forum*, Vol. 71. Pergamon, 114–120.

[46] he Associated Press. 2019. Singer Goo Hara's death shines light on harassment in the cutthroat K-pop industry. https://www.syracuse.com/us-news/2019/11/singer-goo-haras-death-shines-light-on-harassment-in-the-cutthroat-k-pop-industry.html

[47] Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and Brain Sciences* 33, 2-3 (2010), 61–83.

[48] Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. 2017. Deceiving google's perspective api built for detecting toxic comments. *arXiv preprint arXiv:1702.08138* (2017).

[49] Huffpost. 2011. Facebook's censorship problem. https://www.huffpost.com/entry/facebooks-censorship-prob_b_852001

[50] Jane Im, Sarita Schoenebeck, Marilyn Iriarte, Gabriel Grill, Daricia Wilkinson, Amna Batool, Rahaf Alharbi, Audrey Funwie, Tergel Gankhuu, Eric Gilbert, et al. 2022. Women's Perspectives on Harm and Justice after Online Harassment. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–23.

[51] Ronald Inglehart, Christian Haerpfer, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, Bi Puranen, et al. 2014. World values survey: Round six-country-pooled datafile version. *Madrid: JD Systems Institute* (2014), 12.

[52] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. " Did You Suspect the Post Would be Removed?" Understanding User Reactions to Content Removals on Reddit. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–33.

[53] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)* 25, 2 (2018), 1–33.

[54] Jialun'Aaron' Jiang, Skyler Middler, Jed R Brubaker, and Casey Fiesler. 2020. Characterizing community guidelines on social media platforms. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*. 287–291.

[55] Jialun Aaron Jiang, Morgan Klaus Scheuerman, Casey Fiesler, and Jed R Brubaker. 2021. Understanding international perceptions of the severity of harmful content online. *PloS one* 16, 8 (2021), e0256762.

[56] KIM Jinsook. 2021. The Resurgence and Popularization of Feminism in South Korea: Key Issues and Challenges for Contemporary Feminist Activism. *Korea Journal* 61, 4 (2021), 75–101.

[57] Olena Kaminska and Peter Lynn. 2017. Survey-based cross-country comparisons where countries vary in sample design: issues and solutions. *Journal of Official Statistics* 33, 1 (2017), 123–136.

[58] Leslie Kish. 1994. Multipopulation survey designs: five types with seven shared aspects. *International Statistical Review/Revue Internationale de Statistique* (1994), 167–186.

[59] Anti-Defamation League. 2019. More than one-third of Americans experience severe online hate and harassment, new ADL study finds. https://www.adl.org/news/press-releases/more-than-one-third-of-americans-experience-severe-online-hate-and-harassment

[60] Amanda Lenhart, Michele Ybarra, Kathryn Zickuhr, and Myeshia Price-Feeney. 2016. *Online harassment, digital abuse, and cyberstalking in America*. Data and Society Research Institute.

[61] Sebastian Linxen, Christian Sturm, Florian Brühlmann, Vincent Cassau, Klaus Opwis, and Katharina Reinecke. 2021. How weird is CHI?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[62] Sophie Maddocks. 2018. From non-consensual pornography to image-based sexual abuse: Charting the course of a problem with many names. *Australian Feminist Studies* 33, 97 (2018), 345–361.

[63] Carsten Maple, Emma Short, and Antony Brown. 2011. *Cyberstalking in the United Kingdom: An analysis of the ECHO pilot survey*. Technical Report. University of Bedfordshire.

[64] Alice E Marwick. 2021. Morally motivated networked harassment as normative reinforcement. *Social Media+ Society* 7, 2 (2021), 20563051211021378.

[65] J Nathan Matias. 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 (2019), 9785–9789.

[66] Peter Mayer, Yixin Zou, Florian Schaub, and Adam J Aviv. 2021. " Now I'm a bit {angry:}" Individuals' Awareness, Perception, and Responses to Data Breaches that Affected Them. In *30th USENIX Security Symposium (USENIX Security 21)*. 393–410.

[67] Meta. 2022. Principles that guide Meta's stakeholder engagement: Transparency center. (2022). https://transparency.fb.com/policies/improving/principles-guide-our-stakeholder-engagement/

[68] Microsoft. 2002. Digital Civility Index.

[69] Anja Mohorko, Edith de Leeuw, and Joop Hox. 2013. Internet Coverage and Coverage Bias in Europe: Developments Across Countries and Over Time. *Journal of Official Statistics (JOS)* 29, 4 (2013).

[70] Tyler Musgrave, Alia Cummings, and Sarita Schoenebeck. 2022. Experiences of Harm, Healing, and Joy among Black Women and Femmes on Social Media. In *CHI Conference on Human Factors in Computing Systems*. 1–17.

[71] United Nations Department of Economic and Social Affairs. 2019. 2019 Revision of World Population Prospects. (2019).

[72] Lindsey L Osterman and Ryan P Brown. 2011. Culture of honor and violence against the self. *Personality and social psychology bulletin* 37, 12 (2011), 1611–1623.

[73] Sanjana Pampati, Richard Lowry, Megan A Moreno, Catherine N Rasberry, and Riley J Steiner. 2020. Having a sexual photo shared without permission and associated health risks: a snapshot of nonconsensual sexting. *JAMA pediatrics* 174, 6 (2020), 618–619.

[74] Jessica A Pater, Moon K Kim, Elizabeth D Mynatt, and Casey Fiesler. 2016. Characterizations of online harassment: Comparing policies across social media platforms. In *Proceedings of the 19th international conference on supporting group work*. 369–374.

[75] Michael L Pittaro. 2007. Cyber stalking: An analysis of online harassment and intimidation. *International journal of cyber criminology* 1, 2 (2007), 180–197.

[76] Emily Poole. 2015. Fighting back against non-consensual pornography. *USFL Rev.* 49 (2015), 181.

[77] Alison Powell. 2013. Argument-by-technology: How technical activism contributes to Internet governance. In *Research handbook on governance of the Internet*. Edward Elgar Publishing.

[78] Anastasia Powell and Nicola Henry. 2014. Blurred lines? Responding to 'sexting'and gender-based violence among young people. *Children Australia* 39, 2 (2014), 119–124.

[79] Melanie Randall and Vasanthi Venkatesh. 2015. Criminalizing Sexual Violence against Women in Intimate Relationships: State Obligations under Human Rights Law. *American Journal of International Law* 109 (2015), 189–196.

[80] Reuters. 2018. Brother found guilty of murdering Pakistani model Qandeel Baloch in 'honor killing'. https://www.nbcnews.com/news/world/brother-found-guilty-murdering-pakistani-model-qandeel-baloch-honor-killing-n1059431

[81] Sarah T Roberts. 2019. *Behind the screen*. Yale University Press.

[82] Patricia M Rodriguez Mosquera, Agneta H Fischer, Antony SR Manstead, and Ruud Zaalberg. 2008. Attack, disapproval, or withdrawal? The role of honour in anger and shame responses to being insulted. *Cognition and Emotion* 22, 8 (2008), 1471–1498.

[83] Wan Rosalili Wan Rosli, Zaiton Hamin, Ahmad Ridhwan Abd Rani, Saslina Kamaruddin, and Rafizah Abu Hassan. 2021. Non-Criminalisation of Cyber-stalking and Its Impact on Justice for Victims: Some Evidence from Malaysia. *International Journal of Academic Research in Business and Social Sciences* 11, 6 (2021), 1257–1266.

[84] David Ryan. 2018. European remedial coherence in the regulation of non-consensual disclosures of sexual images. *Computer law & security review* 34, 5 (2018), 1053–1076.

[85] Nithya Sambasivan, Amna Batool, Nova Ahmed, Tara Matthews, Kurt Thomas, Laura Sanely Gaytán-Lugo, David Nemer, Elie Bursztein, Elizabeth Churchill, and Sunny Consolvo. 2019. " They Don't Leave Us Alone Anywhere We Go" Gender and Digital Abuse in South Asia. In *proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[86] Willem Saris, Melanie Revilla, Jon A Krosnick, and Eric M Shaeffer. 2010. Comparing questions with agree/disagree response options to questions with construct-specific response options. *Survey Research Methods. 2010; 4 (1): 61-79. DOI: 10.18148/srm/2010. v4i1. 2682* (2010).

[87] Sarita Schoenebeck and Lindsay Blackwell. 2020. Reimagining social media governance: Harm, accountability, and repair. *Yale JL & Tech.* 23 (2020), 113.

[88] Sarita Schoenebeck, Oliver L Haimson, and Lisa Nakamura. 2020. Drawing from justice theories to support targets of online harassment. *new media & society* (2020), 1461444820913122.

[89] Sarita Schoenebeck, Oliver L Haimson, and Lisa Nakamura. 2021. Drawing from justice theories to support targets of online harassment. *new media & society* 23, 5 (2021), 1278–1300.

[90] Sarita Schoenebeck, Carol F Scott, Emma Grace Hurley, Tammy Chang, and Ellen Selkie. 2021. Youth Trust in Social Media Companies and Expectations of Justice: Accountability and Repair after Online Harassment. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–18.

[91] Frances Shaw. 2016. "Bitch I said hi": The Bye Felipe campaign and discursive activism in mobile dating apps. *Social Media+ Society* 2, 4 (2016), 2056305116672889.

[92] Shawna N Smith, Stephen D Fisher, and Anthony Heath. 2011. Opportunities and challenges in the expansion of cross-national survey research. *International Journal of Social Research Methodology* 14, 6 (2011), 485–502.

[93] Jose M Such, Joel Porter, Sören Preibusch, and Adam Joinson. 2017. Photo privacy conflicts in social media: A large-scale empirical study. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 3821–3832.

[94] Sharifa Sultana, Mitrasree Deb, Ananya Bhattacharjee, Shaid Hasan, SM Raihanul Alam, Trishna Chakraborty, Prianka Roy, Samira Fairuz Ahmed, Aparna Moitra, M Ashraful Amin, et al. 2021. 'Unmochon': A Tool to Combat Online Sexual Harassment over Facebook Messenger. (2021).

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

[96] Kea Tijdens and Stephanie Steinmetz. 2016. Is the web a promising tool for data collection in developing countries? An analysis of the sample bias of 10 web and face-to-face surveys from Africa, Asia, and South America. *International Journal of Social Research Methodology* 19, 4 (2016), 461–479.

[97] The World Organization Against Torture. 2004. Violence Against Women in Colombia: A Report to the Committee on Torture. (2004).

[98] The Express Tribune. 2020. Most Pakistanis are in the dark about digital rights, says Nighat Dad. (2020). https://tribune.com.pk/story/2268627/most-pakistanis-are-in-the-dark-about-digital-rights-says-nighat-dad

[99] Jessica Vitak, Kalyani Chadha, Linda Steiner, and Zahra Ashktorab. 2017. Identifying women's experiences with and strategies for mitigating negative effects of online harassment. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 1231–1245.

[100] Laura Vitis and Fairleigh Gilmour. 2017. Dick pics on blast: A woman's resistance to online sexual harassment using humour, art and Instagram. *Crime, media, culture* 13, 3 (2017), 335–355.

[101] Ashley Marie Walker and Michael A DeVito. 2020. "'More gay'fits in better": Intracommunity Power Dynamics and Harms in Online LGBTQ+ Spaces. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

[102] Kate Walker and Emma Sleath. 2017. A systematic review of the current knowledge regarding revenge pornography and non-consensual sharing of sexually explicit media. *Aggression and violent behavior* 36 (2017), 9–24.

[103] wikipedia. 2009. Killing of Neda Agha Soltan. https://en.wikipedia.org/wiki/Killing_of_Neda_Agha-Soltan

[104] Daricia Wilkinson and Bart Knijnenburg. 2022. Many Islands, Many Problems: An Empirical Examination of Online Safety Behaviors in the Caribbean. In *CHI Conference on Human Factors in Computing Systems*. 1–25.

[105] Janis Wolak and David Finkelhor. 2016. Sextortion: Findings from a survey of 1,631 victims. (2016).

[106] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*. 1391–1399.

[107] Sijia Xiao, Coye Cheshire, and Niloufar Salehi. 2022. Sensemaking, Support, Safety, Retribution, Transformation: A Restorative Justice Approach to Understanding Adolescents' Needs for Addressing Online Harm. In *CHI Conference on Human Factors in Computing Systems*. 1–15.

[108] Michele L Ybarra and Kimberly J Mitchell. 2008. How risky are social networking sites? A comparison of places online where youth sexual solicitation and harassment occurs. *Pediatrics* 121, 2 (2008), e350–e357.

[109] Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian D Davison, April Kontostathis, and Lynne Edwards. 2009. Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB* 2 (2009), 1–7.

[110] Jillian C York. 2021. *Silicon Values: The Future of Free Speech Under Surveillance Capitalism*. Verso Books.