Women's Perspectives on Harm and Justice after Online Harassment

JANE IM, University of Michigan, USA
SARITA SCHOENEBECK, University of Michigan, USA
MARILYN IRIARTE, University of Maryland, USA
GABRIEL GRILL, University of Michigan, USA
DARICIA WILKINSON, Clemson University, USA
AMNA BATOOL, University of Michigan, USA
RAHAF ALHARBI, University of Michigan, USA
AUDREY FUNWIE, University of Michigan, USA
TERGEL GANKHUU, University of Michigan, USA
ERIC GILBERT, University of Michigan, USA
MUSTAFA NASEEM, University of Michigan, USA

Social media platforms aspire to create online experiences where users can participate safely and equitably. However, women around the world experience widespread online harassment, including insults, stalking, aggression, threats, and non-consensual sharing of sexual photos. This article describes women's perceptions of harm associated with online harassment and preferred platform responses to that harm. We conducted a survey in 14 geographic regions around the world (N = 3,993), focusing on regions whose perspectives have been insufficiently elevated in social media governance decisions (e.g. Mongolia, Cameroon). Results show that, on average, women perceive greater harm associated with online harassment than men, especially for non-consensual image sharing. Women also prefer most platform responses compared to men, especially removing content and banning users; however, women are less favorable towards payment as a response. Addressing global gender-based violence online requires understanding how women experience online harms and how they wish for it to be addressed. This is especially important given that the people who build and govern technology are not typically those who are most likely to experience online harms.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: online harassment and abuse; gender; social media; online governance

ACM Reference Format:
Jane Im, Sarita Schoenebeck, Marilyn Iriarte, Gabriel Grill, Daricia Wilkinson, Amna Batool, Rahaf Alharbi, Audrey Funwie, Tergel Gankhuu, Eric Gilbert, and Mustafa Naseem. 2022. Women's Perspectives on Harm and Justice after Online Harassment. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 355 (November 2022), 23 pages. https://doi.org/10.1145/3555775

Authors' addresses: Jane Im, University of Michigan, USA; Sarita Schoenebeck, University of Michigan, USA; Marilyn Iriarte, University of Maryland, USA; Gabriel Grill, University of Michigan, USA; Daricia Wilkinson, Clemson University, USA; Amna Batool, University of Michigan, USA; Rahaf Alharbi, University of Michigan, USA; Audrey Funwie, University of Michigan, USA; Tergel Gankhuu, University of Michigan, USA; Eric Gilbert, University of Michigan, USA; Mustafa Naseem, University of Michigan, USA.

2573-0142/2022/11-ART355 $15.00
https://doi.org/10.1145/3555775

1 INTRODUCTION

Online violence against women is a global problem. Women are subjected to abuse, harassment, stalking, threats, and misogyny on social media, and these experiences are increasingly documented in regions around the world [41]. A study by Amnesty International across 8 countries suggests that 23% of women have experienced online harassment, and among those, 41% felt that their physical safety was threatened [4]. A study by UNESCO similarly indicated that 73% of women had experienced or observed some form of online violence [1]. Human rights-based organizations including the United Nations, UNESCO, Human Rights Watch, and others have called for action to address and reduce online violence against women [5, 11, 12, 107].

Yet, social media companies have struggled to effectively govern online harassment. Governance on prominent platforms, such as WeChat, TikTok, and Instagram, relies on a complex and evolving set of policies to determine what content violates community guidelines and what does not [34, 51, 94]. These policies are enacted via a combination of algorithms and human content moderators who detect and process harmful content. However, these processes are imperfect and harmful content can remain on the platform while appropriate content is incorrectly removed [34, 92]. Further, governance processes have insufficiently attended to the disproportionate abuse experienced by marginalized and oppressed groups, such as women, people of color, sex workers, dissidents, transgender people, and disabled people [13, 73, 74, 96, 104, 108, 110].

One reason for these disparities is that platforms embrace principles of fairness that seek to minimize bias in both algorithmic and human moderation processes [34]; however, treating content and users as equal entities falsely presumes that their underlying identities and experiences are equitable [42]. Instead, existing inequities like racism, misogyny, and xenophobia are perpetuated and magnified online, disproportionately harming those groups [23, 53, 62, 77]. While the systematic and structural inequalities experienced online long predate social media, the rise of social media platforms has enabled those inequalities to spread [83, 90, 106, 113]. It is not surprising then, that women continue to experience gender-based harassment including stalking, abuse, misogyny, unsolicited sexual photos ("dick pics"), and non-consensual sexual photo sharing ("revenge porn") with little recourse or remedy [22, 109].

This work builds on a growing interest among social media scholars to center the needs of harassment targets rather than the needs of the perpetrators or the platforms [17, 19, 27, 47, 101, 115]. Current models of governance typically rely on removing content that violates guidelines or temporarily or permanently banning perpetrators [20, 50]. However, scholars have critiqued companies' use of these models because they overlook the experiences of those who are targeted by the harassment, forgoing an opportunity for them to receive just processes or outcomes [101]. Our work aims to understand how platform responses and governance models should be designed to center those who are most affected by online harms.

Further, while gender-based violence is documented in nearly every region and culture around the world—and increasingly online as well—platform governance and policy-making has been largely conducted by relatively few Western perspectives, and these policies are then applied to regions and communities that are vastly different [13, 116]. This work builds on a growing chorus of scholars and activists across communities who are bringing voice and advocacy to women's experiences online outside of Western perspectives (e.g. [66, 95, 96, 105, 116]). Though focusing on women's experiences was not the primary goal at the outset of the project, our analyses indicated that differences between men and women's responses were strong predictors of harms and responses so we focused on those differences. Thus, this article is motivated by the following goals:

- To understand perception of harm associated with different types of online harassment

- To understand preferences for platform responses associated with different types of online harassment
- To identify similarities or differences between women's and men's perceptions and preferences associated with online harassment

We conducted an online survey with nearly 4,000 participants across 14 regions around the world, with a focus on non-Western regions that are underrepresented in social media scholarship. In the survey, participants were presented four different online harassment scenarios and seven possible platform responses. Our analysis shows that perceptions of harm differ by gender, with women perceiving greater harm than men in almost all scenarios. We also find differences in preferences for platform responses—across harassment scenarios, women tend to prefer responses more than men, especially banning accounts, revealing identity of perpetrators, and labeling content. We discuss the complexity of theorizing and applying justice models in the context of gender-based harassment online and conclude with implications for social media platform design and governance.

2 RELATED WORK

2.1 Online Harassment and Harms

Online harassment refers to a wide range of behaviors that are hosted and enabled by technology platforms. These include hate speech, non-consensual sharing of sexual photos, stalking, doxxing, name calling and insults, impersonation, and public shaming [17, 70]. While online harassment was historically depicted as an outlier or fringe behavior, abusive behavior has been endemic in online communities since their inception. Young adults, women, queer people, people of color, disabled people, and other groups are disproportionately affected by online harassment, particularly when those identities intersect [32, 49, 116]. In the U.S., scholars like Jessie Daniels, Lisa Nakamura, Danielle Citron, and others have traced the persistence of racism and misogyny in online spaces for decades, linking those behaviors to underlying offline social experiences [23, 62, 77].

The effects of harassment vary and can include anxiety, stress, fear, humiliation, self-blame, anger, and illness. Online harassment in particular can cause long-term damage due to the persistence and searchability of content. It also has a chilling effect on future disclosures; one study in the United States found that 27% of Internet users were self-censoring their online posts due to fear of harassment [70]. At an extreme, voices are silenced through threats of, or actual, violence and death. Qandeel Baloch, a social media icon in Pakistan, received abuse and harassment for her online persona that resulted in her brother murdering her as an "honor killing" [91]. In South Korea, where the rise of misogyny is recognized as a social issue [53], celebrities Hara Goo and Sulli (Jin-ri Choi) died by suicide after experiencing large-scale cyberbullying and online harassment [40].

Although harassment is instantiated online, targets of online harassment frequently report disruptions to their offline lives, including emotional and physical distress, changes to technology use or privacy behaviors, and increased safety and privacy concerns [28]. Some types of online harassment aim to disrupt a target's offline life, such as swatting (i.e., reporting a crime to induce law enforcement agencies to investigate a target's home) [15]. Targets often choose to temporarily or permanently abstain from social media sites, despite the resulting isolation from information resources and support networks [35]. Online harassment can also be disruptive to personal responsibilities, work obligations, and sleep due to the labor of reporting harassment to social media platforms or monitoring accounts for activity [38, 87].

Though there are emerging efforts to categorize online harms broadly [52], there are not yet standard frameworks for measuring harms associated with online harassment. Harms vary by domain (e.g., medical, environmental), by type of harm, and by severity of harm. The United Nations Declaration on the Elimination of Violence against Women identifies three overarching categories of harm: physical, sexual, and psychological [31]. These three categories are used widely across disciplines and industries and can intersect with other types of harm. Physical harm involves direct bodily injury, or changes in environments that can relate to bodily safety, such as lack of access to housing. Sexual harm relates to sexual abuse, including rape. Psychological harm involves mental and emotional states, and is likely to co-occur with the other two categories; that is, it is unlikely a person could experience physical or sexual harm without also experiencing psychological harm. Other types of harm include financial harm, economic harm, and reproductive harm, though these are not the focus of this paper. Relational harm refers to interpersonal harm experienced by two people or a group of people, such as a family. Family harm can vary from parental neglect to spousal abuse to isolation and abandonment. In the case of online harassment, it can manifest as shame from family members, a phenomenon that has been documented in South Asian regions and may occur in other regions as well [78, 104, 105, 117].

2.2 Platform Responses to Online Harassment

Social media companies have come under increasing criticism for their failures to effectively govern online content [34, 92]. Companies maintain policies to determine what content can or cannot be on their platforms, and enforce those policies through a combination of algorithms and human workers. However, algorithms are crude hammers applied to complex social conflicts; they cannot understand the nuances of social conversations and contexts. As a result, harmful content persists and sometimes even thrives while innocuous content is incorrectly removed [37, 44]. Algorithmic regulation may also magnify harms against marginalized groups (e.g. by removing content that is combating racism while leaving up racist content). Although many of the inequities experienced online are not unique to the Internet, they can be magnified and exacerbated by social media features like "likes", follows, and algorithmic news feeds [71]. Content moderation is also done by human workers who are typically outsourced third-party contractors. These workers are paid low wages for rapid review of content that can be traumatic to look at [92].

Social media companies have often adopted a "neutral" approach to governance that effectively absolves them of the responsibility to adjudicate harm [34]. However, neutrality is not possible when platforms arbitrate millions of personal, nuanced, and contextualized posts daily. Additionally, content moderation decisions are not transparent to users, allowing sites to disguise the power they wield over the process [33, 92]. While most people feel social media companies have a responsibility to remove offensive content from their platforms, few have confidence in companies to determine what offensive content should be removed [67].

One proposal for moving forward is to expand beyond only content removal to consider other kinds of remedies [26, 36, 101]. Sultana et al. explore the idea of a shame-based model of justice for women in the global south, noting that it is necessary in the absence of social and political support for women [104]. In the United States, Schoenebeck et al. [101] explore whether punitive or restorative approaches justice are desired among social media users. Hasinoff, Gibson, and Salehi have proposed that restorative justice approaches focused on repairing harms rather than punishment can improve content moderation [39]. Though not explicitly oriented around justice frameworks, Patel et al. describe user-based (e.g. self-control, self-care) and platform-based (e.g., voting systems, reward systems) responses for detecting toxicity and promoting positivity on platforms [84].

Building on these collective directions, we explore what responses to online harassment are desirable for participants. We frame approaches like removing content or banning users as punitive models that echo criminal justice systems based on removal from society [101]. We also focus on public shaming, such as publicly revealing a perpetrator's identity, given its prominence on social media today [61, 93, 101, 104]. Restorative justice models advocate for repairing harm so we ask survey participants about apologies, while concepts like reparation and racial justice can relate to compensation, so we ask about payment. Our intent is not to be comprehensive, which is unrealistic, but to imagine a range of possibilities for justice frameworks online.

2.3 Online Harassment and Gender

As Internet access has become available in many regions throughout the world, government, NGO, and industry statistics suggest that online abuse—often based on identity—is pervasive. In Mexico, women receive more online abuse than men including defamatory messages and contact through fake accounts [88]. In Mongolia, the majority of the population uses social media daily and although research is limited, hate speech and harassment appear to be widespread [25]. Online harassment is observed in Cameroon, too, despite lower social media adoption rates, and women have protested about online and offline sexual harassment they experience [3, 58]. In South Korea, a massive online sex trafficking on Telegram called the "Nth room case" was discovered in 2020, which included minors' photos being sent to a quarter million users [54, 97, 118]. The Digital Rights Foundation in Pakistan launched a Cyber Harassment Helpline as part of its efforts to support online freedom of expression and the right to privacy for "women, minorities and dissidents" [2].

Despite this global prevalence, Facebook itself has indicated that it focuses on priority areas and areas of high prevalence first and foremost [111]; while it took about 7 days to address employee-reported inauthentic behavior on Facebook for content in the United States, it took 360 days to do the same for content in Mexico. In her book, *Silicon Values*, Jillian York critiques social media platforms for their "signals of mainstream, centralized American media, whose interests lie first and foremost in US affairs—and not just US affairs, but the things that matter most for the country's elites, who are still overwhelmingly white" [116]. Indeed, most regulatory conversations involve an elite few leaders of Silicon Valley or west-coast-based companies who have outsized power over social media as well as its regulation [92]. York explains how the United States public finally became concerned about the power of a concentrated few during the Trump presidency, a concern that has been well-known by activists and scholars globally for a decade or more [116].

This current work is inspired and motivated by efforts to move from a universal perspective to a more "pluriversal" perspective [112]. As authors of a CSCW workshop on decolonizing argue, the pluriversal perspective seeks "to foster 'a world of many worlds' where contradicting ontologies and epistemologies can co-exist without needing to align with each other or claiming more validity over others" [112]. Our work does not fit into decolonialist frameworks which place an emphasis in the geo-political as well as the body-political orientations when conducting research [10]. Nevertheless, our work aims to contribute to the goals of "de-centering" dominant assumptions about who governs social media and how they do so.

The current work builds on movements that have been contesting the US-centered narrative [29, 65, 95, 96, 104, 105]. Sambasivan et al. note that women in South Asian countries report experiences of cyberstalking, impersonation, and personal content leakages that cause emotional harm, reputation harm, romantic coercion, and domestic violence [96]. Women in South Asian regions tend to seek help from friends and family rather than formal channels, though family may not support targets of harassment [104, 105]. Jiang et al. [52] researched how participants from eight countries with the most Facebook users perceive various types of harm associated with online behavior more broadly (e.g., drug use; mutilation). They found that sexualization of minors was highest in harm across countries while spam was lowest, and patterns varied by country. In the domain of content moderation, extensive non-Western work is driven by scholar-activists such as Dia Kayyali, Rasha Abdulla, Nanjira Sambuli, and many others who have raised the alarm about the colossal power of technology companies to decide what speech is allowed or not [116]. Of most concern, they note, are the risks that power brings for freedom of expression, as well as for documentation of human rights violations.

3 METHODS

To explore online harassment harm and responses, we designed an online survey and recruited participants from 14 regions around the world. This section describes survey design and translation, participant recruitment, participants demographics, and data cleaning and analysis.

3.1 Survey Design

We designed the online survey iteratively as a team over a roughly four-month period in 2020. During the survey design phase, we discussed the survey design and brainstormed questions and topics in English. In some cases where multiple members of the research team spoke a shared, non-English language (e.g., Spanish, Urdu), they discussed the questions in their primary languages. We developed and revised the survey numerous times as members of the research team evaluated whether the questions would make sense in the culture of each region being studied. For the regions where the survey was translated into other languages, the research team also evaluated whether the questions still made sense in those languages. Eleven of the 14 regions were represented by members of the research team during the survey design phase; research collaborators from Austria, the Caribbean, and Saudi Arabia were added later. Where questions or wording did not make sense, we iteratively redesigned questions while checking that updated versions would continue to work in the other regions.

To aim for robust translation processes, we hired online translation services (with human translators) to translate the survey from English to the languages used in the surveys. A member of the research team who was bilingual in both the language used in the survey and in English also conducted an independent translation, and then members of the research team compared the independent translations with the paid translations to develop a single version. Each member of the research team then pilot tested the survey with a convenience sample of 2-4 people (typically family and friends) who spoke the relevant language. We used those pilot tests to refine and check translations. We administered the survey in the dominant local language of that region (see Table 2). In India the survey was available in English and Hindi. In Cameroon the survey was available in English because our research team member was from Anglophone Cameroon (not Francophone Cameroon).

To develop the survey questions, we brainstormed varied types of online harassment and their contours—the severity of the harassment, who is affected, and the longevity of the harassment. We started with the six harassment types in Online Harassment Reports from Pew Research (e.g. "Has someone try to purposefully embarrass you") [28] as well as a broad literature review spanning work on content moderation (e.g., [102]), non-consensual image sharing (e.g. [22, 35]), non-Western gendered experiences (e.g. [96, 105]), and other areas. We selected a subset of harassment scenarios based on diversity of type of harassment, variance in severity of harm, and likelihood of being broadly relevant globally. We also focused on harassment scenarios that may have uncertain and inconsistent conceptualizations. For example, while non-consensual image sharing may seem obviously wrong to some people, others have claimed that if a person consensually took sexual photos, it implies them also consenting to those photos to be shared online [22]. Our final survey used four harassment scenarios to minimize participant fatigue.

In the first part of the survey, participants were presented with the the four harassment scenarios (see Table 1): Imagine a person has 1) taken sexual photos of you without your permission and shared them on social media (*sexual photos*); 2) spread malicious rumors about you on social media (*spread rumors*); 3) created fake accounts and sent you malicious comments through direct messages

Harassment scenario	"Imagine a person has [...]"
sexual photos	“taken sexual photos of you without your permission and shared them on social media.”
spread rumors	“spread malicious rumors about you on social media.”
malicious messages	“created fake accounts and sent you malicious comments through direct messages on social media.”
insulted or disrespected	“insulted or disrespected you on social media.”
Perceived harms	
psychological harm	“Would you be concerned for your psychological wellbeing?”
personal safety	“Would you be concerned for your personal safety?”
family reputation	“Would you be concerned for your family reputation?”
sexual harassment	“Would you consider this sexual harassment against you?”
Platform responses	"The social media sites responds by [...]"
removing	“removing the content from the site.”
labeling	“labeling the content as a violation of the site’s rules.”
banning	“banning the person from the site.”
paying	“paying you money.”
apology	“requiring a public apology from the person.”
revealing	“revealing the person’s real name and photograph publicly on the site.”
rating	“by giving a negative rating to the person.”

Table 1. Harassment scenarios, perceived harm, and platform responses.

on social media (*malicious messages*); and 4) insulted or disrespected you on social media (*insulted or disrespected*).

We adapted measures from prior literature to develop four measures associated with harm. For each scenario, participants were asked whether they would be concerned about the following: psychological harm, personal safety, family reputation, and whether they considered the scenario to be sexual harassment. In these adaptations we prioritized interpretability across languages, where everyday people would be likely to understand what we were asking. Thus, they deviate from English-language measures. Specifically, we used “sexual harassment” to capture the sexual nature of the harm, but chose not to use the term “sexual harm” because our pilot testing indicated participants were confused by translations of sexual harm across multiple languages. Sexual harassment is a broad term, like sexual harm, and can refer to both physical and psychological sexual behaviors. Similarly, we used the term “personal safety” as an alternative to physical harm, which also did not translate readily to other languages in our pilot testing.

Perceived harm options were presented on a scale of “Not at all concerned” (coded as 1) to “Extremely concerned” (coded as 5) for psychological harm, personal safety, and family reputation and “Definitely not” (1) to “Definitely” (5) for sexual harassment. The sexual harassment item had a different set of options because it did not pair well with the concerned anchors (“Would you be concerned for sexual harassment” and variations did not make sense). We chose the Not at all concerned/Extremely concerned choice because it translated consistently across languages and because we could put it in the stem of the question to minimize the likelihood of acquiescence bias associated with Strongly Agree/Disagree measures [98].

Then, participants were asked how desirable they found different types of responses to each scenario (see Table 1). To develop the items for the platform responses, we similarly conducted a literature of content moderation practices and policies with a focus on proposed remedies [39, 101, 104]. We chose banning, removing, and labeling as they are prominent examples of existing platform responses [34]. We chose paying, apology, and revealing as they are proposed responses developed in prior work [101]. We added rating because it has been discussed extensively in online communities literature [21, 64] as a compelling approach to regulating online behavior. As with the harassment and harm measures, these are a sampling of possible responses but they are not comprehensive or representative.

The final section contained social media use and demographic questions. The demographic questions were derived from Wave 6 of the World Values Survey (WVS) [48] with some adaptations for use in an online survey (e.g., mobile usability). We chose the WVS because it allowed us to ask single item questions with benchmarks in many countries and languages, which most validated scales do not provide. The WVS is a long survey; we selected questions that captured demographic variables relevant to our project focus. This paper reports on questions related to gender (gender identity, marital status, and number of children) and gender equality (responses to: “When jobs are scarce, men should have more rights to a job than women” and ”When a mother works for pay, the children suffer.”).

The question about participant gender identity expanded the WVS survey version (which includes "Male" and "Female" as options) to also include "Prefer to not disclose" and "Prefer to self-describe." This is aligned with recommendations to move beyond binary options [100] though many scholars also recommend using "Man" and "Woman" instead of "Male" and "Female" which WVS and our survey did not do. Our survey did not ask about non-binary identity because in some countries participants cannot safely identify as a gender outside of male or female. The survey also did not ask about transgender identity or sexual identity, a limitation we return to in the discussion. Scheuerman et al., citing Meissner and Whyte, note that regions around the world have long histories during which gender was not binary which have been overwritten by racism and colonialism [76].

3.2 Participant Recruitment and Demographics

We conducted an online survey with adult participants ages 18 and over from March 2020 through January 2021 across 22 countries: Austria, Cameroon, China, Colombia, India, South Korea, Malaysia, Mexico, Mongolia, Pakistan, Russia, Saudi Arabia, the United States, and nine countries within the Caribbean: St. Kitts and Nevis, Barbados, Dominica, St. Lucia, St. Vincent and the Grenadines, Jamaica, Grenada, Montserrat, and Antigua. This study was exempted from review by our institution’s Institutional Review Board. All participants completed a consent form. We used the survey company Cint to administer the survey in most of the regions in our sample. In the United States, we used the online recruitment platform Prolific. In two regions where Cint (and most global survey companies) have no presence (Caribbean countries, Mongolia), we recruited participants via word of mouth and snowball sampling. The Caribbean and Mongolia samples are convenience samples and will be biased towards who was likely to see our research team members’ invitations to participate. Cint and Prolific use a variety of guardrails to try to ensure diverse and robust survey participants; however, this is also a sample that will be biased towards Internet users and people who are likely to be on survey panels.

Participants were compensated for their time through the survey company or directly, adjusted for exchange rates within the country and for time taken during a pilot test (e.g., Cameroon participants took longer than expected during the pilot survey so we increased their compensation). For the two regions where there was no panel presence, we compensated participants via mobile phone transfer in the local currency (Mongolian Tögrög; East Caribbean Dollar). We aimed to pay a fair wage (e.g., $15 USD/hour in the United States; 2000-3000 Tögrög/hour in Mongolia).

3.2.1 Participant Demographics. Women and men participated in similar rates across regions except for Caribbean countries (women: 69%, men: 27%, see Table 2). The mean age was typically

Region	Language	Number of participants	Age	Men	Women	Prefer not to disclose	Prefer to self-describe
Austria	German	251	37	50%	48%	0%	2%
Cameroon	English	263	26	52%	45%	2%	1%
Caribbean	English	254	27	27%	69%	4%	0%
China	Mandarin	283	36	50%	50%	0%	0%
Colombia	Spanish (Colombian)	296	34	47%	53%	0%	0%
India	Hindi/English	277	32	57%	43%	0%	0%
South Korea	Korean	252	41.5	47%	51%	2%	0%
Malaysia	Malay	298	34	47%	51%	1%	0%
Mexico	Spanish (Mexican)	306	33	51%	49%	0%	0%
Mongolia	Mongolian	367	21	32%	59%	8%	1%
Pakistan	Urdu	302	30	48%	50%	2%	0%
Russia	Russian	282	37	49%	50%	0%	0%
Saudi Arabia	Arabic	258	33	55%	44%	2%	0%
USA	English	304	44	48%	51%	1%	1%

Table 2. Regions studied, survey language, number of participants per region, average age, participant gender ratios.

in the 30s, though Mongolia's median was 21 while South Korea and United States' median were each 41.5 and 44 (Table 2). This pattern skews young but roughly reflects each country 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 the United Nations' population estimates [82]. Participants' self-reported income also diverged across regions, with participants in Austria reporting higher incomes and participants in Caribbean countries self-reporting lower incomes. More than half of the participants had education equivalent to a Bachelor degree for eight regions (Cameroon, China, Colombia, India, Malaysia, Russia, Saudi Arabia, United States); the other regions did not. Participants placed their political views as more "left" (1) than "right" (7) (mean of 3.22 with means ranging across countries from 2.8 in Austria to 3.9 in Korea; participants in Malaysia or Saudi Arabia did not receive this question because they may not be able to safely respond to it).

3.3 Covid Impact Statement

This study was conducted during the beginning of the global coronavirus pandemic. The pandemic inevitably impacted many or all of our participants' lives given its global presence. We do not know how it may have affected their experiences or attitudes about online harassment. To benchmark our sample, we compared mean responses from our participants to mean responses from the World Values Survey. Because our sample and the WVS sample are different (e.g., recruited via online panels with questions optimized for mobile; recruited via door-to-door interviews with verbal question and answer choices) and the comparison is not the focus of the study, we qualitatively report differences. Our benchmarking was conducted with WVS Waves 6 and 7 (we selected Wave 7 when possible since it is more recent, but Wave 6 when questions or response choices were better aligned) for countries that are available in WVS data (China, Colombia, partial India, South Korea, Malaysia, Mexico, Pakistan, Russia, United States).

Participants in our study reported better health than WVS participants. Participants in our study tended to have responses that were more aligned with gender equality for women compared to WVS participants—our participants more strongly disagreed that men should have more rights to a job than women, and disagreed that when a mother works for pay children suffer. This could be due to differences in people who are online versus those who are offline or due to response bias in our survey. It may also be because our sample trended about 5-10 years younger than the WVS samples (and than world population estimates). Our sample was much more likely to have spent money than to have saved it compared to WVS, which could relate to economic downturns in the first six months of Covid (when most data was collected), though it may also be that people who participate in online panels for compensation are looking for income and less likely to be in a position to save money. Participants' responses generally reflected expected trends from WVS data given known social, economic, and political differences (e.g., Pakistan and Saudi Arabia participants tended to have more conservative views about women working than other regions).

3.4 Data Cleaning and Analysis

Data Cleaning. We discarded low-quality responses based on duration of participation (using quantitative thresholds), quality of open-ended question responses (using subjective assessments of quality), and the number of unanswered multiple choice questions (i.e. more than 5 multiple choice questions skipped). Table 2 shows the final number of participants per region after data cleaning. We recruited participants from multiple Caribbean countries (Antigua and Barbuda, Barbados, Dominica, Grenada, Jamaica, Monserrat, St. Kitts and Nevis, St. Lucia, and St. Vincent). While each country of course has its own politics, culture, and economies, we felt that shared experiences across borders justified combining them for analysis. This was also a practical decision; Caribbean countries are small and we wanted to have a similar total sample size to other regions. One coauthor is from one of the Caribbean countries; the other countries are not represented in our research team.

Data Analysis. We analyzed data using R software. We compared means between groups to report women's ratings of harm and justice-seeking responses. Then, we ran linear regressions to compare differences between women and men and to examine other demographic variables. We used the Benjamini–Hochberg (BH) test to correct for multiple comparisons [18]. For the means between groups, we ran Levene's tests to measure variance. For all cases, Levene's tests were significant (i.e., p-values were less than 0.05) indicating that homogeneity of variance assumption is violated. Thus, we used Welch one-way tests (with no assumption of equal variances) for nonparametric data and posthoc pairwise t-tests. The Welch one-way test can be appropriate when there is a sufficiently large sample size [30].

3.5 Positionality Statement

The project team included faculty, graduate students, and undergraduate students based at three universities in the United States. Despite the focus on centering non-Western views, the project is primarily centered at one institution in the United States and coauthors have affiliations with United States universities. Participation in the project included hourly paid research assistant positions, coauthorship, or both. Some project team members were active in the project through the survey design, data collection, and data analysis phases; others were active for only portions of the project. In our interest in decentering Western perspectives, we chose to collect data only in countries where a project team member could speak–as just one voice–to the culture and values of that country. While recognizing that one person does not reflect a country, for this project we decided to focus on regions that the research team collectively shared personal experiences with. As a result, the project team includes people from every country represented in the dataset. By "from," we mean that they have a strong cultural affiliation with the country, including being born in and living in it throughout their childhood and/or early adulthood years. Many of the project team members were physically present in those countries during the project.

4 RESULTS

Results are presented in two sections: perceptions of harm associated with online harassment and preferred responses to online harassment. In each section, we present one visual representation of group means by gender. We use Welch tests to describe means among all participants, then use linear regressions to describe differences by gender and gender-related variables.

4.1 Perceptions of Harm of Online Harassment

We conducted one-way Welch tests to compare means between the four harassment scenarios and the four harm measures. We compared group means between harassment scenarios and again between harm measures. Comparing between harassment scenarios tells us what types of harassment are perceived as more or less harmful (e.g. are rumors more harmful than insults?); comparing between harm types tell us what kinds of harassment might lead to that harm (e.g. is psychological harm or physical harm more prominent?).

Results were significant when comparing means across harassment scenarios for each harm measure: psychological harm, $F(3, 8845.5) = 806.82$, $p < 2.2e\text{-}16$; personal safety, $F(3, 8848.2) = 538.58$, $p < 2.2e\text{-}16$; family reputation, $F(3,8824.1) = 709.52$, $p < 2.2e\text{-}16$; and sexual harassment, $F(3, 8784.7) = 1321$, $p < 2.2e\text{-}16$. They were also significant when comparing means across harm measures for each harassment scenario: sexual photos, $F(3, 8841.7) = 29.967$, $p < 2.2e\text{-}16$; spread rumors, $F(3, 8848.3) = 110.6$, $p < 2.2e\text{-}16$; malicious messages, $F(3, 8846.8) = 29.578$, $p < 2.2e\text{-}16$; insulted or disrespected, $F(3, 8845.3) = 30.175$, $p < 2.2e\text{-}16$.

The *sexual photos* scenario was rated as the most harmful scenario for all types of perceived harm, followed by *spread rumors*, *malicious messages*, and *insulted or disrespected* scenarios (see patterns in Figure 1). Harm measures varied, with sexual photos being rated highest as sexual harassment, spreading rumors being rated highest for family reputation, malicious messages being rated highest for personal safety, and insults or disrespect also rated highest for family reputation.

The post-hoc pairwise tests show that harassment scenarios were mostly different from each other in perceived harm (BH-adjusted $p < .05$), except for sexual harassment in *malicious messages* and *spread rumors* scenarios which were not significantly different from each other (BH-adjusted $p = 0.11$). This is also seen in Figure 1, which shows participants perceived a similar level of sexual harassment from the *malicious messages* and *spread rumors* scenarios.

Similarly, harm measures were mostly different from each other, with differences observed in five out of the six comparisons for each harassment scenario. Exceptions that were not significantly different from each other were personal safety and psychological harm for *spread rumors*, family

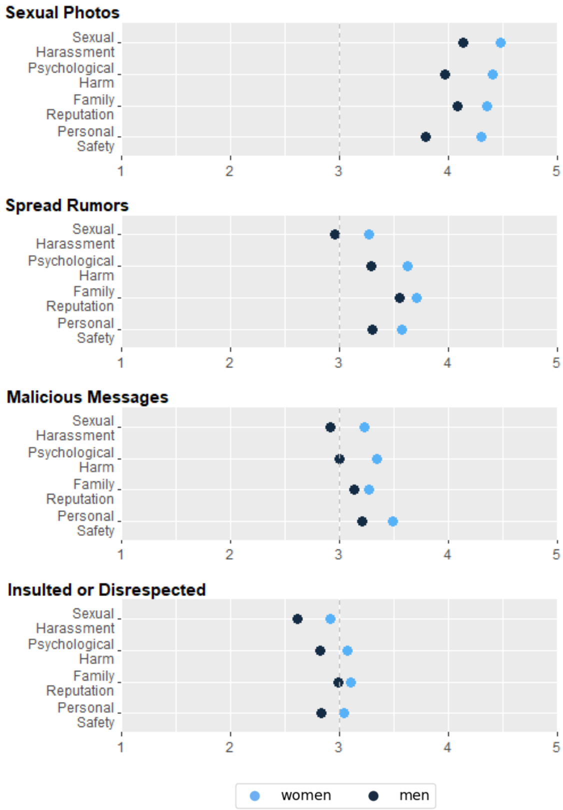

Fig. 1. Perceptions of harm by harassment scenario and harm type. Dark blue represents men's mean and light blue represents women's. 5 indicates the higher rating and 1 indicates lower.

reputation and psychological harm for the *sexual photos*, personal safety and psychological harm for *insulted or disrespected*, and family reputation and psychological harm for *malicious messages*.

4.1.1 Comparing women's perceptions of harm to men's. We fitted a series of linear regression models modeling harm as the dependent variable and demographic variables included in the survey as the independent variables (see Table 3, see visual representation in Figure 1). We ran 16 models for the four harassment scenario and four harm type pairings. Variance inflation factors (VIF) were all less than 2 indicating multicollinearity was not an issue.

Women perceived significantly higher harm than men. This pattern was observed in all 16 harassment scenario-harm pairings with gender accounting for relatively large variability in the models. Results for undisclosed or self-described gender were not significant for most scenario and harm pairings, however, these categories were underrepresented in our data (Table 2). For the *insulted or disrespected* scenario, being married was a predictor of increases in perception of harm for all four measures but this was not observed for other harassment types. For the *malicious*

Sexual Photos	**Psychological Harm**	**Personal Safety**	**Family Reputation**	**Sexual Harassment**
Intercept	4.24 [4.10; 4.38]***	3.80 [3.65; 3.94]***	4.08 [3.94; 4.22]***	4.81 [4.67; 4.94]***
Woman	0.41 [0.34; 0.48]***	0.51 [0.43; 0.58]***	0.26 [0.18; 0.33]***	0.29 [0.22; 0.36]***
Gender undisclosed	0.42 [0.16; 0.69]**	0.70 [0.42; 0.98]***	0.14 [−0.12; 0.41]	−0.02 [−0.28; 0.24]
Gender self-described	−0.35 [−0.93; 0.24]	0.35 [−0.23; 0.93]	−0.11 [−0.67; 0.46]	−0.37 [−0.91; 0.18]
Women jobs	−0.13 [−0.22; −0.05]**	−0.09 [−0.18; 0.01]	−0.04 [−0.12; 0.05]	−0.37 [−0.45; −0.28]***
Mother works	−0.05 [−0.13; 0.04]	0.09 [0.01; 0.18]*	0.01 [−0.07; 0.10]	−0.11 [−0.19; −0.03]**
Is married	−0.00 [−0.09; 0.08]	−0.08 [−0.16; 0.01]	0.07 [−0.02; 0.15]	−0.07 [−0.15; 0.01]
Has children	−0.01 [−0.04; 0.02]	0.03 [−0.00; 0.06]	0.02 [−0.01; 0.05]	0.02 [−0.01; 0.05]
Adj. R^2	0.04	0.05	0.013	0.053
Spread Rumors				
Intercept	3.13 [2.97; 3.29]***	3.03 [2.87; 3.19]***	3.13 [2.97; 3.29]***	2.76 [2.60; 2.92]***
Woman	0.35 [0.27; 0.43]***	0.27 [0.19; 0.36]***	0.18 [0.10; 0.27]***	0.32 [0.24; 0.41]***
Gender undisclosed	0.39 [0.08; 0.70]*	0.29 [−0.02; 0.60]	0.12 [−0.19; 0.43]	0.26 [−0.04; 0.57]
Gender self-described	0.46 [−0.19; 1.10]	0.03 [−0.63; 0.68]	0.02 [−0.63; 0.68]	−0.18 [−0.84; 0.48]
Women jobs	0.07 [−0.03; 0.17]	0.10 [−0.00; 0.20]	0.18 [0.08; 0.28]***	−0.01 [−0.11; 0.09]
Mother works	0.01 [−0.09; 0.10]	0.04 [−0.06; 0.13]	0.05 [−0.05; 0.14]	0.08 [−0.01; 0.18]
Is married	0.12 [0.02; 0.21]*	0.04 [−0.06; 0.13]	0.08 [−0.01; 0.18]	0.20 [0.11; 0.30]***
Has children	0.00 [−0.03; 0.04]	0.08 [0.05; 0.12]***	0.07 [0.03; 0.10]***	0.02 [−0.02; 0.06]
Adj. R^2	.019	.018	.016	.022
Malicious Messages				
Intercept	2.80 [2.63; 2.97]***	3.13 [2.96; 3.29]***	2.60 [2.42; 2.78]***	2.82 [2.66; 2.98]***
Woman	0.35 [0.26; 0.43]***	0.26 [0.17; 0.35]***	0.17 [0.08; 0.27]***	0.30 [0.21; 0.38]***
Gender undisclosed	0.31 [−0.01; 0.63]	0.21 [−0.11; 0.53]	0.27 [−0.06; 0.61]	0.12 [−0.19; 0.43]
Gender self-described	−0.54 [−1.22; 0.13]	−0.47 [−1.14; 0.21]	−0.62 [−1.32; 0.09]	−0.33 [−0.99; 0.32]
Women jobs	0.06 [−0.05; 0.16]	−0.03 [−0.13; 0.08]	0.16 [0.05; 0.27]**	−0.09 [−0.20; 0.01]
Mother works	0.03 [−0.07; 0.13]	0.03 [−0.07; 0.13]	0.10 [−0.00; 0.21]	0.09 [−0.00; 0.19]
Is married	0.03 [−0.07; 0.13]	−0.03 [−0.13; 0.07]	0.07 [−0.03; 0.18]	0.10 [0.01; 0.20]*
Has children	0.08 [0.04; 0.11]***	0.10 [0.06; 0.14]***	0.13 [0.09; 0.17]***	0.07 [0.03; 0.10]***
Adj. R^2	0.021	0.017	0.026	0.021
Insult or Disrespect				
Intercept	2.23 [2.06; 2.40]***	2.23 [2.05; 2.40]***	2.13 [1.95; 2.30]***	2.27 [2.11; 2.43]***
Woman	0.31 [0.22; 0.40]***	0.26 [0.17; 0.35]***	0.18 [0.09; 0.28]***	0.34 [0.25; 0.42]***
Gender undisclosed	0.23 [−0.09; 0.55]	0.43 [0.10; 0.76]*	0.37 [0.03; 0.71]*	0.14 [−0.17; 0.45]
Gender self-described	−0.53 [−1.20; 0.14]	−0.09 [−0.79; 0.60]	−0.16 [−0.88; 0.55]	0.34 [−0.31; 0.99]
Women jobs	0.29 [0.19; 0.40]***	0.25 [0.14; 0.36]***	0.36 [0.25; 0.47]***	0.05 [−0.05; 0.15]
Mother works	0.07 [−0.03; 0.17]	0.09 [−0.01; 0.20]	0.17 [0.06; 0.27]**	0.11 [0.01; 0.20]*
Is married	0.29 [0.19; 0.39]***	0.20 [0.10; 0.31]***	0.20 [0.09; 0.30]***	0.24 [0.15; 0.34]***
Has children	−0.02 [−0.06; 0.02]	0.05 [0.02; 0.09]**	0.06 [0.02; 0.10]**	0.02 [−0.02; 0.05]
Adj. R^2	0.033	0.03	0.036	0.027

Table 3. Linear regression models for perceptions of harm showing coefficients and confidence intervals.

messages scenario, having children was a predictor of increases in perception of harm for all four measures. This may be because sending comments through direct messages on social media is a fear parents often have for their children.

4.1.2 Attitudes about gender and perceptions of harm. For three of the harassment scenarios (*spread rumors*, *insulted or disrespected*, *malicious messages*), participants who tended to agree with the statement "When jobs are scarce, men should have more rights to a job than women" perceived greater harm to family reputation. This suggests that people who are less likely to support gender equality are also more concerned about family reputation after online harassment. In contrast, for the *sexual photos* scenario, people who tended to disagree with the statement rated those scenarios as higher in terms of psychological harm and sexual harassment.

4.2 Preferred Responses to Online Harassment

In the prior section, we measured perceptions of harm for each of the four harassment scenarios. Here we want to understand which responses are preferred for each of the harassment scenarios. This allows us to answer the question, if someone experiences that type of online harassment, what response might they prefer? In the last section, we compared group means between harassment scenarios (one comparison for each of the four harm measures) and also between the harm measures (one comparison for each of the four harassment scenarios) because it was useful to be able to interpret both. Here we only compare group means between response types (for the four harassment scenarios) because it is primarily useful to know what responses are preferred for a given type of harassment.

One-way Welch tests for preferred responses were significant for all four harassment scenarios: *sexual photos* scenario, $F(6, 12381) = 304.87$, $p < 2.2e\text{-}16$; *spread rumors* scenario, $F(6, 12395) = 291.24$, $p < 2.2e\text{-}16$; *malicious messages* scenario, $F(6, 12399) = 307.57$, $p < 2.2e\text{-}16$; and *insulted or disrespected* scenario, $F(6, 12402) = 262.18$, $p < 2.2e\text{-}16$.

As evident in Figure 2, removing, banning, and labeling were the three most preferred responses. Payment was the least preferred in all four harassment scenarios. Posthoc pairwise t-tests showed that most ratings of responses differed significantly

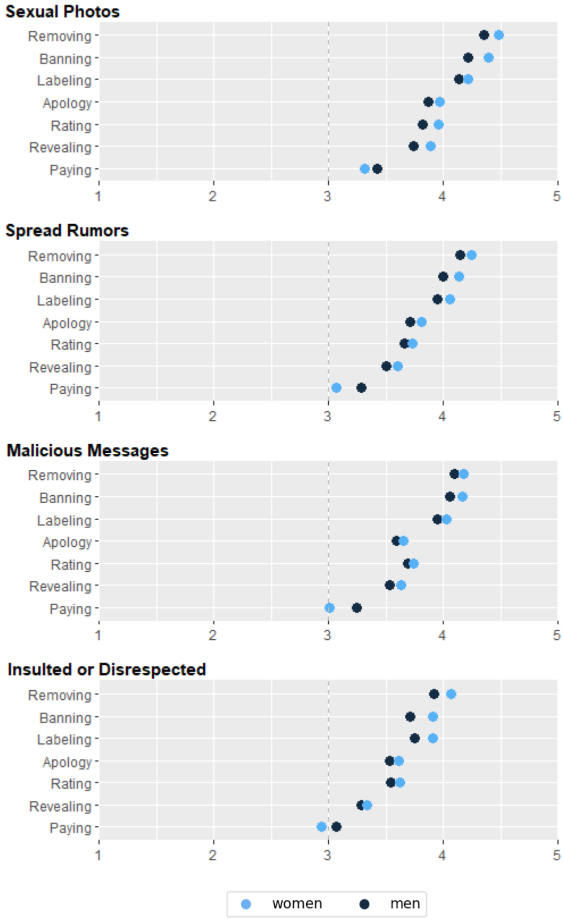

Fig. 2. Preferences for responses by harassment scenario and response type. Dark blue represents men's mean and light blue represents women's. 5 indicates the higher rating and 1 indicates lower.

for most, though not all, pairings. The rating and apology responses tended not to be significantly different from each other.

4.2.1 Comparing women's preferred responses to men's. We fitted a series of linear regression models for each of the four harassment scenarios and each of the seven response options (28 total, see Table 4). Platform response options were modeled as the dependent variable and demographic data related to gender as the independent variables. Multicollinearity was not an issue as VIF were all less than 2.

The difference in preferences between women and men was observed most frequently for banning: women prefer banning more than men in all four harassment scenarios. Women also preferred the apology response compared to men for all four harassment scenarios. This was followed by revealing which was preferred by women more than men in three of the four scenarios. Labeling and rating were preferred by women in two scenarios and removing was preferred in one scenario. However, one outlier is payment which was inverted; women preferred payment *less* than men did in three of the harassment scenarios.

Sexual Photos	**Removing**	**Labeling**	**Banning**	**Payment**	**Apology**	**Revealing**	**Rating**
Intercept	5.05 [4.92; 5.17]***	4.62 [4.47; 4.77]***	4.86 [4.72; 5.00]***	3.57 [3.38; 3.76]***	3.91 [3.74; 4.08]***	3.78 [3.61; 3.96]***	4.02 [3.85; 4.19]***
Woman	.06 [−.01; .12]	.04 [−.04; .11]	.13 [.06; .20]***	−.11 [−.21; −.01]*	.11 [.02; .20]*	.16 [.07; .26]***	.11 [.02; .20]*
Undisclosed	.08 [−.16; .33]	−.14 [−.42; .15]	−.03 [−.29; .24]	−.52 [−.89; −.15]**	.07 [−.26; .40]	.03 [−.30; .37]	.28 [−.05; .62]
Self-described	−.26 [−.77; .25]	−.48 [−1.08; .13]	−.36 [−.91; .20]	−.41 [−1.19; .36]	−.17 [−.86; .51]	−.41 [−1.11; .28]	.15 [−.54; .85]
Women jobs	−.39 [−.47; −.31]***	−.27 [−.37; −.18]***	−.39 [−.48; −.31]***	−.16 [−.28; −.04]**	−.02 [−.12; .09]	−.08 [−.19; .03]	−.19 [−.30; −.09]***
Mother works	−.12 [−.20; −.05]**	−.11 [−.20; −.03]*	−.12 [−.20; −.03]**	−.07 [−.18; .04]	−.09 [−.19; .01]	−.04 [−.14; .07]	−.04 [−.14; .06]
Is married	−.08 [−.16; −.01]*	−.04 [−.13; .05]	.03 [−.05; .11]	.11 [−.01; .22]	.11 [.01; .21]*	.16 [.06; .27]**	−.02 [−.12; .08]
Has children	.03 [.01; .06]*	.04 [.01; .08]**	.02 [−.01; .05]	.08 [.04; .12]***	.02 [−.02; .05]	.03 [−.01; .07]	.09 [.05; .13]***
Adj. R^2	.04	.017	.037	.011	.003	.008	.012

Spread Rumors							
Intercept	4.62 [4.48; 4.75]***	4.32 [4.17; 4.47]***	4.34 [4.19; 4.48]***	3.11 [2.93; 3.29]***	3.51 [3.35; 3.68]***	3.20 [3.03; 3.37]***	3.69 [3.53; 3.86]***
Woman	.07 [−.01; .14]	.08 [.01; .16]*	.12 [.04; .19]**	−.19 [−.28; −.09]***	.12 [.03; .20]**	.12 [.03; .21]*	.06 [−.03; .14]
Undisclosed	−.04 [−.30; .23]	.02 [−.26; .30]	.06 [−.22; .34]	−.20 [−.55; .15]	.40 [.08; .73]*	.18 [−.15; .52]	.27 [−.04; .59]
Self-described	−.50 [−1.05; .06]	−.42 [−1.01; .16]	−.61 [−1.19; −.03]*	−.74 [−1.47; −.01]*	−.62 [−1.29; .05]	−.63 [−1.33; .06]	−.14 [−.80; .53]
Women jobs	−.28 [−.37; −.20]***	−.23 [−.32; −.14]***	−.30 [−.39; −.20]***	−.11 [−.23; −.00]*	.07 [−.03; .18]	.02 [−.09; .12]	−.12 [−.22; −.02]*
Mother works	−.11 [−.19; −.02]*	−.14 [−.23; −.06]**	−.04 [−.12; .05]	.11 [.00; .21]*	−.04 [−.14; .06]	.06 [−.04; .16]	−.04 [−.14; .05]
Is married	−.03 [−.11; .05]	.07 [−.02; .16]	.05 [−.03; .14]	.08 [−.03; .19]	.15 [.05; .25]**	.14 [.04; .24]**	.07 [−.03; .17]
Has children	.05 [.02; .08]***	.07 [.04; .11]***	.06 [.03; .10]***	.10 [.06; .14]***	.04 [.01; .08]*	.11 [.07; .15]***	.11 [.07; .15]***
Adj. R^2	.022	.022	.022	.015	.01	.02	.015

Malicious Messages							
Intercept	4.56 [4.42; 4.70]***	4.34 [4.20; 4.49]***	4.49 [4.34; 4.63]***	3.07 [2.89; 3.26]***	3.33 [3.16; 3.50]***	3.37 [3.20; 3.54]***	3.81 [3.64; 3.97]***
Woman	.02 [−.05; .09]	.06 [−.02; .14]	.09 [.01; .16]*	−.21 [−.30; −.11]***	.09 [.00; .18]*	.11 [.02; .20]*	.03 [−.06; .11]
Undisclosed	−.13 [−.40; .14]	−.10 [−.38; .18]	−.04 [−.32; .23]	−.45 [−.80; −.09]*	.14 [−.19; .47]	−.27 [−.59; .05]	−.03 [−.34; .29]
Self-described	−.46 [−1.03; .11]	−.13 [−.72; .46]	−.44 [−1.01; .13]	−.22 [−.95; .52]	−.74 [−1.42; −.05]*	−.96 [−1.64; −.28]**	−.16 [−.82; .51]
Women jobs	−.29 [−.38; −.20]***	−.22 [−.31; −.13]***	−.27 [−.36; −.18]***	−.10 [−.22; .01]	.05 [−.06; .16]	−.01 [−.11; .10]	−.19 [−.30; −.09]***
Mother works	−.09 [−.18; −.01]*	−.16 [−.25; −.07]***	−.09 [−.17; −.01]*	.11 [−.00; .21]	.02 [−.08; .12]	.01 [−.09; .11]	−.04 [−.13; .06]
Is married	.03 [−.05; .11]	.09 [.00; .18]*	.07 [−.01; .16]	.12 [.01; .23]*	.23 [.13; .33]***	.20 [.10; .30]***	.14 [.04; .23]**
Has children	.05 [.02; .08]**	.05 [.02; .08]**	.02 [−.01; .05]	.07 [.03; .12]***	.03 [−.01; .06]	.06 [.02; .10]**	.09 [.06; .13]***
Adj. R^2	.019	.018	.018	.015	.011	.016	.017

Insulted or Disrespected							
Intercept	4.19 [4.04; 4.34]***	3.94 [3.78; 4.09]***	3.69 [3.54; 3.85]***	2.73 [2.55; 2.92]***	3.28 [3.11; 3.44]***	2.84 [2.66; 3.01]***	3.52 [3.35; 3.68]***
Woman	.12 [.04; .20]**	.15 [.07; .23]***	.21 [.13; .29]***	−.08 [−.17; .02]	.12 [.03; .21]**	.09 [−.00; .18]	.10 [.01; .18]*
Undisclosed	−.16 [−.45; .12]	−.01 [−.30; .29]	.23 [−.07; .53]	−.08 [−.42; .27]	.26 [−.06; .59]	−.03 [−.38; .31]	−.05 [−.37; .26]
Self-described	−.55 [−1.15; .06]	.00 [−.62; .62]	−.39 [−1.02; .24]	−.40 [−1.13; .32]	−.13 [−.81; .55]	−.35 [−1.06; .37]	−.12 [−.78; .55]
Women jobs	−.20 [−.29; −.10]***	−.17 [−.26; −.07]***	−.10 [−.20; −.00]*	.01 [−.11; .12]	.11 [.00; .21]*	.06 [−.06; .17]	−.12 [−.22; −.01]*
Mother works	−.07 [−.16; .02]	−.08 [−.17; .01]	.01 [−.08; .10]	.10 [−.01; .21]	−.03 [−.13; .06]	.11 [.01; .22]*	−.02 [−.12; .08]
Is married	−.00 [−.09; .08]	.08 [−.01; .17]	.16 [.06; .25]***	.15 [.04; .26]**	.25 [.15; .35]***	.27 [.16; .37]***	.12 [.02; .22]*
Has children	.08 [.04; .11]***	.07 [.03; .10]***	.04 [.00; .07]*	.07 [.03; .11]***	.01 [−.03; .05]	.07 [.03; .11]***	.09 [.06; .13]***
Adj. R^2	.016	.015	.014	.012	.011	.022	.015

Table 4. Linear regression models for preferences for responses showing coefficients and confidence intervals.

Participants who tended to agree with the statement "When jobs are scarce, men should have more rights to a job than women" were less likely to prefer responses for 18 out of the 28 pairings. In other words, people who tended not to support gender equality were also less supportive of many responses to online harassment. Interestingly, participants who had children preferred 22 of the 28 responses. This likely means that participants who are parents were thinking about their children when responding to the questions even if the questions were directed to the parent.

4.3 Limitations when interpreting results

This work used surveys to capture people's perceptions on online harassment and various approaches to harm online. We chose multiple-choice questions as our goal was to compare responses from a large sample of participants and regions. However, future research could use qualitative methods to better understand ideas about harm and conceptualizations of harassment.

While the authors are from different regions represented in the dataset, the project is centered at one institution in the United States. This means our perspectives are influenced by U.S.-centered approaches and values. Researchers based in other regions can contribute valuable perspectives by designing cross-cultural studies grounded in non-Western perspectives and justice frameworks [85].

As mentioned in Section 3.1, we selected a non-representative and non-comprehensive set of harassment scenarios and cannot fully explain variances in harm and response ratings for these scenarios. For example, while taking sexual photos without permission and sharing them on social media clearly elicited higher harm, we do not know how participants interpreted the prompt or what the specific kinds of harms they were imagining were.

Lastly, we acknowledge the use of frequentist statistics can lead to over- or mis-interpretation of results. We focused on prominent themes in our results which may be less prone to arise out of chance.

5 DISCUSSION

5.1 Women Perceive Greater Harm Than Men

Our results show that women perceive greater harm associated with online harassment than men across many regions around the world. The scenario of taking sexual photos without permission and sharing them on social media was highest in harm for both women and men, and was higher for women than men. Advocates and activists globally have been fighting for greater protection of women's rights related to non-consensual sharing of sexual photos. In her book, *Nobody's Victim*, United States-based lawyer Carrie Goldberg writes about non-consensual sharing, "We are putting tech companies on notice. For too long, dating apps and other digital products have been enabling [people] to commit heinous crimes that put us all at risk. It's time these companies are held accountable. They need to think differently about the responsibility they owe us all." [35].

These calls to action are needed across both corporations and policy-making. Scholar Nanjira Sambuli describes how an anti-pornography law in Uganda that could be intended to support victims may also be the law that can be used to punish women for images shared online without their consent [72]. Mariana Valente, based in Brazil, has pointed out that misogynistic cultures in Brazil have led to criminal cases involving online defamation suits by men against women for calling them sexist, rather than cases that protect women [75]. In South Korea, Team Flame, a team of women undergraduate students that publicized the Nth Room case—the cyber-trafficking case—led to greater sanctions for sharing non-consensual videos online and revised laws for social platforms to curb online sexual violence [45, 114]. In the United States, Goldberg, along with legal scholars Danielle Citron and Mary Anne Franks, have been pivotal in pushing through greater regulatory protection to prevent non-consensual sexual image sharing [22, 35, 57].

However, learning that men perceive lower harm than women in all four types of harassment, not just non-consensual sharing of sexual photos, across all regions studied, is concerning. Scholars and technologists may perceive only some kinds of harassment as gendered (e.g. those that are explicitly gendered like sexist comments), but how people experience various interactions may vary based on their prior experiences or identities [56]. Importantly, interactions may not have to contain overt sexual language or expressions to cause harm. Many prominent technology companies are predominantly led by men and there may be a misalignment between how they believe women experience harm on their platforms versus how women see themselves experiencing harm. This misalignment may be especially difficult to detect when harms are difficult to measure; unlike some harms (e.g., theft of a bicycle, a broken arm), it remains difficult to quantify how online harassment harms its targets.

A harm-based perspective may help social media companies to better address disparities on their platform. That is, platforms could adopt guiding frameworks, drawn from human rights and civil rights principles, that seek to recognize and repair harms. Such an approach could also help to address different experiences of harm in different regions. For example, in Colombia there have been widespread concerns and growing public awareness about violence against women. Hundreds of women are murdered per year on account of their gender in Colombia [24] and as a result, Colombian women (and some men) have joined the online movement #NiUnaMenos and #NiUnaMás (#NotOneLess and #NotOneMore) to reflect on unequal power relationships that result in subordination and violence.

These movements highlight how when we talk about online harassment and harm, we must consider the broader contexts in which those harms are taking place. Though harms can be difficult to quantify or measure, they also invite the perspectives of those who experience the harm and could enable us to imagine different kinds of responses to them.

5.2 Exploring Justice Models While Centering Those Most Affected by Harassment

Our results showed that participants are generally favorable towards most of the platform responses. That is, on average, they rate all of the responses as more desirable than undesirable. We intentionally chose a wide range of platform responses that reflect a range of justice frameworks. For example, apologies are aligned with restorative justice frameworks that seek to recognize and repair harm [101], whereas revealing identities publicly invokes a kind of vigilante justice that relies on public shaming [81]. Participants' general favorability towards all of these platform responses, despite the differences between them, suggests that—to some extent—they may be more eager to have *any* kind of remedy rather than nothing.

Women prefer banning more than men in all four harassment scenarios. This may reflect differing frequency or severity of harassment that women experience online. Given that women experience various kinds of gender-based harm throughout their lives [14], banning users could be motivated by a sense of urgency to address and minimize any potential for future harm. Whereas an apology may be appropriate for less severe harassment, banning users is a more severe sanction and may be especially desirable for more severe kinds of online harassment.

Schoenebeck et al. [101] argue that existing models of governance—namely, removing content and banning users—reflect criminal justice models that focus on punishing perpetrators of harassment. Building on arguments for criminal justice reform, they suggest that punitive models may fail to acknowledge the experience of targets or to hold perpetrators accountable for harms. Alternative justice theories, such as restorative justice, racial justice, or even shame as a form of justice may provide new avenues for how platforms respond to harassment [39, 101, 104]. While alternative justice theories could be compelling in some contexts, our findings show that they should be engaged with carefully and contextually when centering the voices of people who are harmed. In particular, our findings of women preferring banning—a punitive measure—more than men indicates women may prefer more strict sanctions, especially in the absence of any other effective remedies. Our results also indicate that though women were generally more favorable than men about most of the platform responses, they were less favorable towards payment as a response.

Though future work should examine why, one possible interpretation is that punitive measures like banning meet victims' needs in ways which current restorative measures do not when their major concern is not reparation [103]. Simultaneously, while both men and women preferred apologies less than banning, women preferred it more than men, indicating that nuanced responses may be needed depending on the nature of the harassment. More research should examine what women's reasons are for these preferences and understanding the dynamics between punitive and restorative measures on social platforms.

As many Internet scholars have argued, context matters. Differences in cultures and judicial systems across regions may impact women's preferences—for example, being a victim of gender-based online harassment might heighten the risk of reputational damage for the victim in regions with strong patriarchal or misogynistic values [6–9, 60]. We also speculate that for some regions, trust in judicial systems may shape people's preferences for responses. In many countries, legal systems are criticized for taking sexual assault cases lightly, making it difficult for people who are harmed to find justice through them [69]. This may cause people to prefer punitive and decisive responses existing on social platforms like removal and banning.

Finally, a number of countries in our sample are post-colonial countries. While it is widely acknowledged that the colonial experience affected the political development of ex-colonial countries [79], it may have also have helped shape punitive online governance preferences, as colonial justice was largely punitive in nature [63, 79].

Ultimately, our results contribute to the conversations around the importance of centering those who are most affected by online harm in social media governance. Importantly, preferences may vary by cultures and individuals, suggesting that one-size-fits-all remedies may be ineffective or harmful in themselves [52, 55, 89, 101, 105]. Though we focused on women's experiences, it is likely that these cumulative experiences of harm shape how targets in a variety of groups—non-binary people, transgender people, people of color in the United States, lower caste people in India, etc—may prefer to seek justice in their online experiences.

5.3 Categories and Universality in Online Harassment

Violence against women is a global cause embraced by the United Nations, World Health Organization, Amnesty International, and other organizations [5, 11, 12, 107]; however, it essentializes gender rather than recognizing the mutability and non-binary properties of gender [16, 59, 99]. Gender binaries are social, economic, and political categories [46, 86]. In some of the countries we studied, it is illegal to be transgender or non-binary, and in many of them, it is not socially or culturally accepted (e.g., [80]). We chose not to ask about non-binary identity or transgender identity because responding truthfully to those questions could put participants at risk in some regions; at the same time, overlooking them further marginalizes those groups.

In our analysis, we also wrestled with the treatment of countries. We chose to analyze responses in aggregate to recognize universality of experiences and by country (plus the collection of countries in the Caribbean) in recognition of the cultural, social, economic, and political conditions of those regions. This was partly a conceptual and analytical choice—"global" or "non-Western" studies often define populations within the boundaries of a country. It was also a methodological one—country is a required selection criteria in panels.

However, recognizing universality is a contested matter. In many ways universality may signify what Gramsci refers to as "cultural hegemony," in which beliefs and value systems are mediated by those in power [68]. Anna Lauren Hoffmann, based in the United States, writes about the epistemologies and ideals associated with binary categories [43]. Hoffmann calls on readers to critically reflect on the idea of universalism which is often conflated with imperialism or the West [43]. Doing this does not mean giving up on universalism itself, but merely that what has been labeled universal is often far from it. In the context of women's experiences online, there may be some near-universal experiences of harm which reflect offline inequalities. However, there are also regional differences which reflect cultural values, policies, and laws that shape women's experiences of justice online. Addressing these requires that people and institutions in positions of power shift from aspirations of neutrality, towards more principled governance that centers the well-being of those who experience systematic harm.

6 CONCLUSION

This work reveals that on average, women perceive higher harm associated with online harassment than men do across 14 regions around the world. Perceived harm is highest for sharing of non-consensual sexual photos. It also reveals that on average, revealing and banning were the most preferred platform responses. When considering gender, women participants prefer banning users and apologies more than men, but they prefer payment less than men. Responses to harassment are generally desirable for all participants; however, women find most responses more desirable than men do. This work contributes to an expanding movement to decenter US perspectives on social media governance, while centering conversations about global and local values. It also draws attention to the limitations of principles like neutrality in content moderation if they are enacted in fundamentally inequitable social contexts.

ACKNOWLEDGMENTS

We thank Anandita Aggarwal, Ting-Wei Chang, Chao-Yuan Cheng, Yoojin Choi, Banesa Hernandez, Kseniya Husak, Jessica Jamaica, Wafa Khan, and Nurfarihah Mirza Mustaheren for their contributions to this project. We thank Michaelanne Thomas, David Nemer, and Katy Pearce for early conversations about these ideas. We thank members of the Social Media Research Lab for their feedback at various stages of the project. We also thank the reviewers for their constructive comments. Im would especially like to thank Yoon Jeong Cha, Yoojin Choi, Song Mi Lee, and Wookjae Maeng for their advice and Shubham Atreja, Agrima Seth, and Anjali Singh for proofreading the paper. This material is based upon work supported by the National Science Foundation under Grants 1763297 and 1552503 and by a gift from Instagram.

REFERENCES

[1] [n.d.]. Combatting Online Violence Against Women & Girls: A Worldwide Wake-up Call. https://en.unesco.org/sites/default/files/highlightdocumentenglish.pdf

[2] [n.d.]. Digital Rights Foundation. https://digitalrightsfoundation.pk/

[3] [n.d.]. #StopSexualHarassment237: Cameroon women wan end work-place harassment.

[4] 2017. Amnesty reveals alarming impact of online abuse against women. https://www.amnesty.org/en/latest/news/2017/11/amnesty-reveals-alarming-impact-of-online-abuse-against-women/

[5] 2020. Take five: Why we should take online violence against women and girls seriously during and beyond COVID-19. https://www.unwomen.org/en/news/stories/2020/7/take-five-cecilia-mwende-maundu-online-violence

[6] Norah Abokhodair, Adam Hodges, and Sarah Vieweg. 2017. Photo sharing in the Arab Gulf: Expressing the collective and autonomous selves. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 696–711.

[7] Lila Abu-Lughod. 2016. *Veiled sentiments: Honor and poetry in a Bedouin society.* Univ of California Press.

[8] Farah Ahmad, Natasha Driver, Mary Jane McNally, and Donna E Stewart. 2009. "Why doesn't she seek help for partner abuse?" An exploratory study with South Asian immigrant women. *Social science & medicine* 69, 4 (2009), 613–622.

[9] Yeslam Al-Saggaf. 2016. An exploratory study of attitudes towards privacy in social media and the threat of blackmail: The views of a group of Saudi women. *The Electronic Journal of Information Systems in Developing Countries* 75, 1 (2016), 1–16.

[10] Syed Mustafa Ali. 2016. A brief introduction to decolonial computing. *XRDS: Crossroads, The ACM Magazine for Students* 22, 4 (2016), 16–21.

[11] Z Aziz. 2015. Eliminating Online Violence Against Women and Engendering Digital Equality. sl: OHCHR. (2015). https://www.ohchr.org/Documents/Issues/Women/WRGS/GenderDigital/DueDiligenceProject.pdf

[12] Heather Barr. 2019. Internet Bringing New Forms of Violence Against Women. https://www.hrw.org/news/2019/10/28/internet-bringing-new-forms-violence-against-women#

[13] Catherine Barwulor, Allison McDonald, Eszter Hargittai, and Elissa M Redmiles. 2021. "Disadvantaged in the American-dominated Internet": Sex, Work, and Technology. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[14] Melanie A Beres. 2007. 'Spontaneous' sexual consent: An analysis of sexual consent literature. *Feminism & Psychology* 17, 1 (2007), 93–108. https://doi.org/10.1177/0959353507072914

[15] Laura-Kate Bernstein. 2016. Investigating and prosecuting swatting crimes. *US Att'ys Bull.* 64 (2016), 51.

[16] Rena Bivens and Oliver L Haimson. 2016. Baking gender into social media design: How platforms shape categories for users and advertisers. *Social Media+ Society* 2, 4 (2016), 2056305116672486.

[17] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 24 (Dec. 2017), 19 pages. https://doi.org/10.1145/3134659

[18] Frank Bretz, Torsten Hothorn, and Peter Westfall. 2016. *Multiple comparisons using R*. Chapman and Hall/CRC.

[19] Annemarie Bridy. 2018. Remediating social media: A layer-conscious approach. *BUJ Sci. & Tech. L.* 24 (2018), 193.

[20] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–22.

[21] Yan Chen, F Maxwell Harper, Joseph Konstan, and Sherry Xin Li. 2010. Social comparisons and contributions to online communities: A field experiment on movielens. *American Economic Review* 100, 4 (2010), 1358–98.

[22] Danielle Keats Citron and Mary Anne Franks. 2014. Criminalizing revenge porn. *Wake Forest L. Rev.* 49 (2014), 345.

[23] Jessie Daniels. 2009. *Cyber racism: White supremacy online and the new attack on civil rights*. Rowman & Littlefield Publishers.

[24] Joe Parkin Daniels. 2021. 'Nowhere is safe': Colombia confronts alarming surge in femicides. https://www.theguardian.com/global-development/2021/jan/25/nowhere-is-safe-colombia-confronts-alarming-surge-in-femicides

[25] Munkhchimeg Davaasharav. 2019. Mongolia to combat hate speech in online media. https://asia.fes.de/news/mongolia-to-combat-hate-speech-in-online-media

[26] Evelyn Douek. 2020. Governing Online Speech: From'Posts-As-Trumps' to Proportionality and Probability. *Columbia Law Review* 121, 1 (2020).

[27] Evelyn Douek. 2020. The Limits of International Law in Content Moderation. *UCI Journal of International, Transnational, and Comparative Law (forthcoming 2021)* (2020).

[28] Maeve Duggan. 2017. Online harassment 2017. (2017). https://ncvc.dspacedirect.org/handle/20.500.11990/10

[29] Michaelanne Dye, Annie Antón, and Amy S Bruckman. 2016. Early adopters of the Internet and social media in Cuba. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*. 1295–1309.

[30] 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.

[31] 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

[32] Global Fund for Women. [n.d.]. Online violence: Just because it's virtual doesn't make it any less real. https://www.globalfundforwomen.org/online-violence-just-because-its-virtual-doesnt-make-it-any-less-real/

[33] Tarleton Gillespie. 2010. The politics of 'platforms'. *New media & society* 12, 3 (2010), 347–364.

[34] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

[35] Carrie Goldberg. 2019. *Nobody's Victim: Fighting Psychos, Stalkers, Pervs, and Trolls*. Plume.

[36] Eric Goldman. 2021. Content Moderation Remedies. *Michigan Technology Law Review, Forthcoming* (2021).

[37] Kishonna L Gray. 2020. *Intersectional Tech: Black users in digital gaming*. LSU Press.

[38] Mark Griffiths. 2002. Occupational health issues concerning Internet use in the workplace. *Work & Stress* 16, 4 (2002), 283–286.

[39] EA Hasinoff, AD Gibson, and N Salekhi. 2020. The Promise of Restorative Justice in Addressing Online Harm. https://www.brookings.edu/techstream/the-promise-of-restorative-justice-in-addressing-online-harm/

[40] 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

[41] Lori Heise. 1994. Gender-based abuse: the global epidemic. *Cadernos de Saúde Pública* 10 (1994), S135–S145.

[42] Anna Lauren Hoffmann. 2019. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society* 22, 7 (2019), 900–915.

[43] Anna Lauren Hoffmann. 2021. Even When You Are a Solution You Are a Problem: An Uncomfortable Reflection on Feminist Data Ethics. *Global Perspectives* 2, 1 (2021).

[44] 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).

[45] Sung hwa Hong. 2020. After the Nth Room: South Korea Combating Digital Sex Crime. https://times.uos.ac.kr/news/articleView.html?idxno=10012

[46] Janet Shibley Hyde, Rebecca S Bigler, Daphna Joel, Charlotte Chucky Tate, and Sari M van Anders. 2019. The future of sex and gender in psychology: Five challenges to the gender binary. *American Psychologist* 74, 2 (2019), 171.

[47] Jane Im, Jill Dimond, Melody Berton, Una Lee, Katherine Mustelier, Mark S. Ackerman, and Eric Gilbert. 2021. Yes: Affirmative Consent as a Theoretical Framework for Understanding and Imagining Social Platforms *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 403, 18 pages. https://doi.org/10.1145/3411764.3445778

[48] 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.

[49] Amnesty International. [n.d.]. Toxic Twitter - A Toxic Place For Women. https://www.amnesty.org/en/latest/research/2018/03/online-violence-against-women-chapter-1/

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] Kim Joohee and Jamie Chang. 2021. Nth Room Incident in the Age of Popular Feminism: A Big Data Analysis. *Azalea: Journal of Korean Literature & Culture* 14, 14 (2021), 261–287.

[55] Naveena Karusala, Apoorva Bhalla, and Neha Kumar. 2019. Privacy, patriarchy, and participation on social media. In *Proceedings of the 2019 on Designing Interactive Systems Conference*. 511–526.

[56] Roger C Katz, Roseann Hannon, and Leslie Whitten. 1996. Effects of gender and situation on the perception of sexual harassment. *Sex Roles* 34, 1-2 (1996), 35–42.

[57] Danielle Keats Citron. 2018. Sexual privacy. *Yale LJ* 128 (2018), 1870.

[58] Simon Kemp. 2019. Digital 2019: Cameroon. https://datareportal.com/reports/digital-2019-cameroon

[59] Os Keyes. 2018. The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on human-computer interaction* 2, CSCW (2018), 1–22.

[60] Adeel Khan and Rafat Hussain. 2008. Violence against women in Pakistan: Perceptions and experiences of domestic violence. *Asian Studies Review* 32, 2 (2008), 239–253.

[61] Kate Klonick. 2015. Re-shaming the debate: Social norms, shame, and regulation in an internet age. *Md. L. Rev.* 75 (2015), 1029.

[62] Beth Kolko, Lisa Nakamura, and Gilbert Rodman. 2013. *Race in cyberspace*. Routledge.

[63] Elizabeth Kolsky. 2010. Colonial justice in British India.

[64] Robert E Kraut and Paul Resnick. 2012. *Building successful online communities: Evidence-based social design*. Mit Press.

[65] Neha Kumar. 2020. Encountering feminisms across borders. *Interactions* 27, 6 (2020), 46–48.

[66] Neha Kumar, Naveena Karusala, Azra Ismail, Marisol Wong-Villacres, and Aditya Vishwanath. 2019. Engaging feminist solidarity for comparative research, design, and practice. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[67] John Laloggia and Inquiries. 2019. US Public has little confidence in social media companies to determine offensive content. *Pew Research Center (July 2019)* (2019).

[68] TJ Jackson Lears. 1985. The concept of cultural hegemony: Problems and possibilities. *The American Historical Review* (1985), 567–593.

[69] William Lee. 2021. (2021).

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

[71] Rebecca Lewis. 2018. Alternative influence: Broadcasting the reactionary right on YouTube. *Data & Society* 18 (2018).

[72] Evelyn Lirri. [n.d.]. The Challenge of Tackling Online Violence Against Women in Africa. https://www.opennetafrica.org/the-challenge-of-tackling-online-violence-against-women-in-africa/

[73] Alice E Marwick and Robyn Caplan. 2018. Drinking male tears: Language, the manosphere, and networked harassment. *Feminist Media Studies* 18, 4 (2018), 543–559.

[74] Allison McDonald, Catherine Barwulor, Michelle L Mazurek, Florian Schaub, and Elissa M Redmiles. 2021. " It's stressful having all these phones": Investigating Sex Workers' Safety Goals, Risks, and Practices Online. In *30th USENIX Security Symposium (USENIX Security 21)*. 375–392.

[75] Darija Medić. [n.d.]. Internet De-Tox: A Fail-Proof Regimen to End Online Sexism. https://dig.watch/sessions/internet-de-tox-fail-proof-regimen-end-online-sexism

[76] Shelbi Nahwilet Meissner and Kyle Powys Whyte. 2017. Theorizing indigeneity, gender, and settler colonialism. *in The Routledge Companion to the Philosophy of Race, eds. Paul Taylor, Lina Alcoff and Luvell Anderson. New York: Routledge* (2017), 152–167.

[77] Lisa Nakamura. 2008. *Digitizing race: Visual cultures of the Internet.* Vol. 23. U of Minnesota Press.

[78] Mustafa Naseem, Fouzia Younas, and Maryam Mustafa. 2020. Designing Digital Safe Spaces for Peer Support and Connectivity in Patriarchal Contexts. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–24.

[79] Mohd. A. Nawawi. 1971. Punitive Colonialism: The Dutch and the Indonesian National Integration. *Journal of Southeast Asian Studies* 2, 2 (1971), 159–168. http://www.jstor.org/stable/20069916

[80] Fayika Farhat Nova, Michael Ann Devito, Pratyasha Saha, Kazi Shohanur Rashid, Shashwata Roy Turzo, Sadia Afrin, and Shion Guha. 2021. "Facebook Promotes More Harassment": Social Media Ecosystem, Skill and Marginalized Hijra Identity in Bangladesh. *arXiv preprint arXiv:2102.02853* (2021).

[81] Martha C Nussbaum. 2009. *Hiding from humanity: Disgust, shame, and the law.* Princeton University Press.

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

[83] Nizan Geslevich Packin. 2020. Disability Discrimination Using AI Systems, Social Media and Digital Platforms: Can We Disable Digital Bias? *Social Media and Digital Platforms: Can We Disable Digital Bias* (2020).

[84] Aashka Patel, Christine L Cook, and Donghee Yvette Wohn. 2021. User Opinions on Effective Strategies Against Social Media Toxicity. In *Proceedings of the 54th Hawaii International Conference on System Sciences.* 3005.

[85] Sachin R Pendse, Amit Sharma, Aditya Vashistha, Munmun De Choudhury, and Neha Kumar. 2021. "Can I Not Be Suicidal on a Sunday?": Understanding Technology-Mediated Pathways to Mental Health Support. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* 1–16.

[86] V Spike Peterson. 1999. Political identities/nationalism as heterosexism. *International Feminist Journal of Politics* 1, 1 (1999), 34–65.

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

[88] Martha Pskowski. [n.d.]. Mexican women stand up to cyberattacks and vicious digital violence. https://www.pri.org/stories/2017-08-24/mexican-women-stand-cyberattacks-and-vicious-digital-violence

[89] Hawra Rabaan, Alyson L Young, and Lynn Dombrowski. 2021. Daughters of Men: Saudi Women's Sociotechnical Agency Practices in Addressing Domestic Abuse. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–31.

[90] Kunal Relia, Zhengyi Li, Stephanie H Cook, and Rumi Chunara. 2019. Race, ethnicity and national origin-based discrimination in social media and hate crimes across 100 US cities. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 417–427.

[91] Reuters. [n.d.]. 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

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

[93] Jon Ronson. 2016. *So you've been publicly shamed.* Riverhead Books.

[94] Lotus Ruan, Jeffrey Knockel, Jason Q Ng, and Masashi Crete-Nishihata. 2016. One App, Two Systems: How WeChat uses one censorship policy in China and another internationally. (2016).

[95] Nithya Sambasivan, Nova Ahmed, Amna Batool, Elie Bursztein, Elizabeth Churchill, Laura Sanely Gaytan-Lugo, Tara Matthews, David Nemar, Kurt Thomas, and Sunny Consolvo. 2019. Toward Gender-Equitable Privacy and Security in South Asia. *IEEE Security & Privacy* 17, 4 (2019), 71–77.

[96] 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.

[97] Choe Sang-Hun. 2021. South Korean Man Gets 34 Years for Running Sexual Exploitation Chat Room. https://www.nytimes.com/2021/04/08/world/asia/korea-sex-crime-chat-rooms.html

[98] 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).

[99] Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. 2019. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–33.

[100] Morgan Klaus Scheuerman, Katta Spiel, Oliver L Haimson, Foad Hamidi, and Stacy M Branham. 2020. HCI guidelines for gender equity and inclusivity. *UMBC Faculty Collection* (2020).

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

[102] Joseph Seering, Felicia Ng, Zheng Yao, and Geoff Kaufman. 2018. Applications of social identity theory to research and design in social computing. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW'18). ACM, New York, NY, USA.*

[103] Julie Stubbs. 2007. Beyond apology? Domestic violence and critical questions for restorative justice. *Criminology & Criminal Justice* 7, 2 (2007), 169–187.

[104] 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).

[105] Sharifa Sultana, François Guimbretière, Phoebe Sengers, and Nicola Dell. 2018. Design within a patriarchal society: Opportunities and challenges in designing for rural women in bangladesh. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[106] Brendesha M Tynes, Henry A Willis, Ashley M Stewart, and Matthew W Hamilton. 2019. Race-related traumatic events online and mental health among adolescents of color. *Journal of Adolescent Health* 65, 3 (2019), 371–377.

[107] UNESCO. 2021. UNESCO calls to end online violence against women journalists in 8 March campaign. https://en.unesco.org/news/unesco-calls-end-online-violence-against-women-journalists-8-march-campaign

[108] 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.

[109] 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.

[110] 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.

[111] Julia Carrie Wong. [n.d.]. Revealed: the Facebook loophole that lets world leaders deceive and harass their citizens. https://www.theguardian.com/technology/2021/apr/12/facebook-loophole-state-backed-manipulation

[112] Marisol Wong-Villacres, Adriana Alvarado Garcia, Juan F Maestre, Pedro Reynolds-Cuéllar, Heloisa Candello, Marilyn Iriarte, and Carl DiSalvo. 2020. Decolonizing Learning Spaces for Sociotechnical Research and Design. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*. 519–526.

[113] Yue Wu, Yi Mou, Yi Wang, and David Atkin. 2018. Exploring the de-stigmatizing effect of social media on homosexuality in China: An interpersonal-mediated contact versus parasocial-mediated contact perspective. *Asian Journal of Communication* 28, 1 (2018), 20–37.

[114] So-Yeon Yoon. 2020. The spark that ignited the 'Nth room' fire. https://koreajoongangdaily.joins.com/2020/03/31/features/The-spark-that-ignited-the-Nth-room-fire/3075527.html

[115] Jillian York. [n.d.]. THE SANTA CLARA PRINCIPLES On Transparency and Accountability in Content Moderation. https://santaclaraprinciples.org/

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

[117] Fouzia Younas, Mustafa Naseem, and Maryam Mustafa. 2020. Patriarchy and social media: Women only facebook groups as safe spaces for support seeking in Pakistan. In *Proceedings of the 2020 International Conference on Information*

and Communication Technologies and Development. 1–11.

[118] Ji-Yeong Yun. 2020. Feminist Net-Activism as a New Type of Actor-Network that Creates Feminist Citizenship. *Asian Women* 36, 4 (2020).

Received April 2021; revised November 2021; accepted March 2022