Drawing from Justice Theories to Support Targets of Online Harassment

Sarita Schoenebeck[1], Oliver Haimson[1], Lisa Nakamura[2]
School of Information[1], American Culture Department[2]
University of Michigan
{yardi,haimson,lnakamur}@umich.edu

Abstract

Most content moderation approaches in the U.S. rely on criminal justice models that sanction offenders via content removal or user bans. However, these models write the online harassment *targets* out of the justice-seeking process. Via an online survey with U.S. participants (*N*=573), this research draws from justice theories to investigate approaches for supporting targets of online harassment. We uncover preferences for banning offenders, removing content, and apologies but aversion to mediation and adjusting targets' audiences. Preferences vary by identities (e.g., transgender participants on average find more exposure to be undesirable; American Indian or Alaska Native participants on average find payment to be unfair) and by social media behaviors (e.g., Instagram users report payment as just and fair). Our results suggest that a one-size-fits-all approach will fail some users while privileging others. We propose a broader theoretical and empirical landscape for supporting online harassment targets.

Introduction

Social media sites have developed a set of complex processes for responding to online harassment (Pater, Kim, Mynatt, & Fiesler, 2016). These processes, which are largely developed within U.S.-based companies and cultures, focus on determining whether content violates community guidelines, and if so, whether and how to sanction offenders who have violated those guidelines. For example, content that discriminates against another person or group may be removed and the offender may be warned or outright banned (Bradford et al., 2019). On the other hand, content may be offensive to some users but deemed not in violation of community guidelines and thus left in place (Gillespie, 2018; Roberts, 2019). Regardless of outcome, targets of offensive or harmful content receive little or no notification during the content moderation process, preventing them from experiencing acknowledgement or reparation of the harms they may have experienced. Indeed, "processes optimized solely for stopping harassment are unlikely to address the larger impact of the harassment on the targeted user" (Matias et al., 2015). Further, people from protected social groups (e.g., based on gender, sex, race, religion, or disability) are more likely to be targets of harassment on social media, perpetuating and magnifying injustices they experience in their lives (Duggan, 2017). These individuals tend to be undercompensated for their online labor (Postigo, 2016) and expend disproportionate work to be included online (Ahmed, 2012).

Content moderation approaches mirror principles from criminal justice systems in the U.S., which focus on punishing offenders rather than restoring justice to victims (Bobo & Thompson, 2006; Cole, 1999; Wenzel, Okimoto, Feather, & Platow, 2008). Criminal justice theories propose that crime should be met with proportionate punishment (e.g., ranging from fines to imprisonment). As social media sites have grown dramatically in scale, they have adopted criminal justice approaches to regulation where people who violate rules or norms are warned or removed from the community (Matias et al., 2015; Pater et al., 2016). This work builds on a growing movement in the U.S. that recognizes the criminal justice system's limitations in supporting *targets* of an offense. While other studies have examined harassment perpetrators' (Munger, 2017) and harassment reporters' (Matias et al., 2015) experiences, this study was designed to prioritize the voices of the targets themselves. This research critically examines social media sites' responses to online harassment and lays a path for integrating justice into the governance process.

Online Governance

Online governance models rely on a combination of policies, norms, tools, administrators, workers, and computation. Governance models can be either top-down (in which harassment policies and moderation are imposed by the platform), bottom-up (in which users decide on and impose their own rules and moderation strategies), or a combination of the two (Bradford et al., 2019). Top-down approaches to moderation typically seek to position sites as neutral, displaying standardized guidelines that "perform, and therefore reveal in oblique ways, how platforms see themselves as public arbiters of cultural value" (Bradford et al., 2019; Caplan, 2018; Pater et al., 2016; Gillespie, 2018). However, content moderation decisions are invisible to users, allowing sites to disguise the power they wield over the process (Gillespie, 2010; Roberts, 2019). 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 (Laloggia & Inquiries, 2019). In contrast, bottom-up approaches rely on volunteer moderation practices that require extensive uncompensated labor from volunteers (Postigo, 2016). On Reddit and Twitch, for example, subreddits or channels rely on volunteer moderators to establish and enforce site policies and norms (Matias, 2019; Wohn, 2019, Seering et al, 2019), and strategies for handling harassment include educating, sympathizing, shaming, humor, and blocking (Cai and Wohn, 2019). Volunteer moderation is often buttressed by automated bots or systems (e.g., Chandrasekharan et al, 2019) that support moderation demands at scale. Moderation is also performed by users themselves, either individually via report options or en masse (e.g., third-party blocklists (Jhaver et al, 2018)).

This work aims to radically reconsider how social media sites should support targets of online harassment. We focus on top-down governance to encourage broader and more equitable governance practices from companies who have a responsibility to support online harassment targets. Drawing from justice theories, our goal is to uncover approaches that recognize power differentials and are responsive to people with a wide range of abilities, identities, and preferences. Such approaches acknowledge that designing for people without power requires designing for everybody, by seeking to eradicate those systems of power ("The Combahee River Collective Statement," n.d.). While regulatory approaches can increase social media companies' responsibility to their users (e.g., rights to privacy, right to not be discriminated against), such approaches may not inspire trust or confidence from targets of online harassment given that U.S. laws and criminal justice systems have historically been complicit with institutions like slavery and sexism (Bobo & Thompson, 2006; Cole, 1999).

Theories of Justice and Online Responses

On social media, targets of online harassment have few opportunities to experience visibility or reparation. Even if offensive content is removed from the site or an offender is banned from the site, the target can experience harms that feel isolating and invisible. While the criminal justice system has a strong foothold in U.S. justice systems (Cole, 1999; Wenzel et al., 2008), the past few decades have seen increased interest in theories of justice, including restorative justice, racial justice, and social justice, that prioritize rehabilitation and reparation rather than punishment (Bell, 2008; Jackson, 2013; Wenzel et al., 2008).

Restorative justice is concerned with mediation processes that mend conflict between an offender, a victim, and the community, often with the involvement of facilitators (Braithwaite, 1999; Wenzel et al., 2008; Zehr, 2015). Restorative justice requires that offenders acknowledge wrongdoing, accept responsibility for their actions, and express remorse, typically via an apology. However, apologies currently play little role in criminal justice procedures in the U.S. Bibas and Bierschbach (2004) note, "Our criminal justice system works as a speedy assembly line: It plea bargains cases efficiently and maximizes punishment for the limited resources available. This assembly line leaves little room for remorse and apology". Additionally, apologies have been discouraged in legal proceedings because they may invoke admissions of responsibility or blameworthiness (Scher & Darley, 1997).

Theories of racial and economic justice acknowledge the systematic injustices and inequities communities have experienced via lack of access to employment, education, housing, and other rights (Bell, 2008; Cole, 1999; Fallon & Weiler, 1984; Jackson, 2013). Most views towards racial justice advocate for deliberate systems and support to promote racial equity rather than simply removing discrimination. Ahmed (2012) notes, "Describing the problem of racism can mean being treated as if you have created the problem, as if the very talk about division is what is divisive". Indeed, critical race theory grew out of the need to understand the differential forms of oppression that apply to multiple identities such as race, class, and gender: laws were not designed to treat everyone equally, and any just application of the law must acknowledge this and work to align laws to this reality. Proponents of reparations have called attention to the generational downstream effects of slaves' inability to earn wages for their work or acquire literacy and education (Coates, 2014; Nelson, 2016). They also have called attention to long legacies of trauma and grief across generations as a result of displacement and genocide (Brave Heart & DeBruyn, 1998). An economic justice approach argues that people should be given what allows them to lead a fruitful life, including payment for work that is done (Jackson, 2013).

Racial and economic justice overlap with the concept of social justice, which Rawls popularized as the distribution of benefits and burdens across individuals and social groups (Rawls, 2009). More contemporary lens have critiqued Rawls' "veil of ignorance" approach which overlooks individual identities and experiences. Instead, social justice scholars argue identity should be central to interpretations of justice, which acknowledge identity as complex and fluid (e.g., what it means to be an immigrant shifts based on how immigrants are treated) (Clayton & Opotow, 2003). Further, social justice advocates for spaces for individuals to participate safely within a shared set of values.

Measuring Justice

Justice has sometimes been conflated with the concept of fairness, and is often measured using the language of fairness in surveys. For example, distributive justice considers whether benefits are distributed fairly across individuals and is typically measured with questions about fairness (Sunshine & Tyler, 2003; Tyler, 1994). Procedural justice (which is sometimes referred to as procedural fairness) considers whether processes were perceived as fair, independent of outcome (Sunshine & Tyler, 2003; Tyler, 1994). Consistent with procedural justice theory, increased transparency in content moderation explanations increases perception that a decision was fair (Jhaver et al, 2019). Facebook is working to increase procedural justice in its content moderation, however, it is largely focused on how to increase perceptions of fairness among offenders (Bradford et al., 2019). In contrast, our work seeks to promote justice in outcomes for harassment targets.

Justice acknowledges structural power differentials and seeks to dismantle them, whereas fairness maintains power differentials, because it locates the source of problems within individuals or technologies (D'Ignazio & Klein, 2018). Hoffmann argues that the concept of fairness falsely attributes harm to individuals instead of systemic and contextual problems (Hoffmann, 2019). That argument has been recognized by computer scientists, who acknowledge the limitations of fairness in mathematical representations because of the values and politics they embed (Barocas, Hardt, & Narayanan, n.d.). Fairness further falls short because racism, sexism, and various forms of discrimination are fundamentally different from other kinds of rule violations.

Thus, while social media sites' enforcement of rules about copyright violation and appropriate use may enact justice in appropriate ways, when these same enforcement mechanisms are used to address harmful behaviors like hate speech, users feel unseen, unheard, and doubly harmed (Citron, 2014). Hoffmann and Jonas (2016) call for a more expansive notion of justice that exposes how technology companies' power can create hostile environments for vulnerable or otherwise disadvantaged populations, with little legal or regulatory oversight (Citron, 2014).

Our survey measured three variables—justice, fairness, and desirability—of responses to online harassment. We developed these measures by first conducting a preliminary study to understand participants' perceptions of justice and fairness in response to online harassment. Our preliminary study paired an online experiment with a free response survey, and found that participants perceived banning or blocking users as just responses, and banning users or legal regulation as fair responses. It also showed that favorability towards justice and fairness of social media governance can vary by identity, a theme the current study develops further. In terms of desirability, what is desirable may not necessarily align with what is just or fair, and vice versa. Rawls (2009) proposed that justice principles should order a society stably over time because people will develop a desire to act in accordance with those principles. In practice, of course, society does not converge so easily. In the context of social media regulation, *we define fairness as the correct enforcement of previously-stated rules that a user has violated*. In contrast, *we view justice as the effective remediation of harms arising from interpersonal conflict, prejudice, and harassment*. Justice centers the experiences and perspective of the person who suffered harm. Finally, *we define desirability as simply, what users find desirable as an outcome.*

Hypotheses and Research Questions

Traditional approaches to responding to online harassment (i.e., removing content, banning users) are the status quo, but people may value alternative approaches that enact justice in more holistic and comprehensive ways (e.g., apology, mediation, payment). Thus, we ask:

RQ1: Are there differences in attitudes towards perceived a) justice, b) fairness, and/or c) desirability of traditional versus alternative actions taken by social media sites?

Traditional criminal justice systems in the U.S. have perpetuated discrimination and inequities towards people from marginalized identities, and these injustices are perpetuated online. We investigate whether people from non-dominant identities (e.g., racial minorities, transgender people) prefer alternative approaches to the criminal justice model, and conversely, whether people from dominant identities prefer the status quo.

H1a: Non-dominant social groups will be more favorable towards perceived a) justice, b) fairness, and/or c) desirability of alternative actions taken by social media sites, as opposed to traditional actions.

And the corollary:

H1b: Dominant social groups will be more favorable towards perceived a) justice, b) fairness, and/or c) desirability of traditional actions taken by social media sites, as opposed to alternative actions.

Identity attributes (e.g., race, gender, political views, socioeconomic class, and age) have all influenced how people experience justice through social, technological, and legal systems in the U.S. Here, we examine the relationship between identity and attitudes towards how social media sites should respond to online harassment.

RQ2 - 6: Is [race/ethnicity [RQ2] / gender [RQ3] / political orientation [RQ4] / socioeconomic class [RQ5] / age [RQ6]] associated with attitudes towards perceived a) justice, b) fairness, and/or c) desirability of actions taken by social media sites?

Social media sites have different rules and norms that govern appropriate behavior. Here we ask whether frequency of use on six different major platforms influences attitudes towards how social media sites should respond to harassment.

RQ7: Is frequency of use associated with attitudes towards perceived a) justice, b) fairness, and/or c) desirability of actions taken by social media sites?

Social media sites evaluate whether content was in violation of community guidelines in isolation, which prevents them from recognizing sustained harassment over time (Blackwell, Dimond, Schoenebeck, & Lampe, 2017; Duggan, 2017; Massanari, 2017). Here we examine whether past experiences being targets or perpetrators of harassment, or supporters or targets of harassment, influences attitudes towards how social media sites should respond to that harassment.

RQ8: Are prior experiences of being harassed or harassing others on social media associated with attitudes towards perceived a) justice, b) fairness, and/or c) desirability of actions taken by social media sites?

Finally, orientations towards justice could influence preferences for actions sites take. Some actions might be perceived as more fair in terms of process (procedural justice) or outcome (distributive justice).

RQ9: Are orientations towards procedural or distributive justice associated with attitudes towards perceived a) justice, b) fairness, and/or c) desirability of actions taken by social media sites?

Overview of Methods

This study was deemed exempt by our institution's ethics review board. Participants completed an online consent form. We pre-registered research questions and hypotheses on the Open Science Foundation[1] before data collection began (RQ9 was added after).

We used the language of "aggressive or hostile on social media" as proxies for "online harassment." Both terms are susceptible to response bias, but "online harassment" has been used in mainstream media in socially and politically charged contexts and may be more susceptible to a variety of biases. We developed the alternative language via a focus group then pilot tested the language iteratively. We use "aggressive or hostile" when referring to our instruments, and related language like harassment, bad behavior, and abuse when describing general phenomena. We chose not to define justice, fairness, or desirability because they cannot be concisely defined in a survey item, and because we wanted to elicit participants' favorability to those concepts based on their own internal interpretations. As a result, participants' responses will reflect their own varied conceptualizations of what justice and fairness mean. We pre- tested the anchor "restore justice to me" with 30 online participants to confirm that they were able to interpret the phrase.

Survey development

We conducted an online, anonymous survey of adults in the U.S. We developed ten items iteratively via brainstorming and discussions among the research team and pilot testing multiple times. The final survey contained two items that represent traditional actions on major platforms and the remaining eight items were novel, alternative actions not currently used by major platforms. The traditional actions reflect criminal justice theories while the alternative actions reflect a range of values embedded across criminal, racial, economic, and social justice theories. Each question contained the stem: "Imagine that a person is being aggressive or hostile to you on social media. The social media site responds by [options in Table 1]." Each of the ten items was presented as a bipolar matrix with three rows measuring justice, fairness, and desirability.

Action	Survey prompt

[1] Preregistration: https://osf.io/ja4d8/?view_only=1bccdbff142a4257b4f92cf4d9032713

Traditional social media site actions	
REMOVING CONTENT	"removing the content from the site"
BANNING USERS	"banning the person from the site"
Alternative social media site actions	
PAYMENT	"paying you and your supporters"
APOLOGY	"requiring a public apology from the person"
OFFENDER LIST	"adding the person to an online public list of offenders"
MEDIATION	"facilitating an online meeting including you, the person, and a mediator to discuss your experience"
IDENTITY	"educating the person about your identities and experiences"
LESS EXPOSURE	"allowing you to be less exposed to a wide audience on the site"
MORE EXPOSURE	"allowing you to have more exposure to a large audience on the site"
OWN SPACE	"empowering you to have a space with your own rules and values"

Table 1. Potential social media site responses to online harassment.

The next section asked about prior experiences being perpetrators, targets, or supporters of targets of online harassment. We adapted scales from Sunshine & Tyler (2003) to measure participants' beliefs about procedural justice (three items) and distributive justice (three items) among social media sites. Internal consistency of the procedural justice scale was reliable with α=.84. Distributive justice was not reliable with α=.33, possibly because two of the questions were difficult to interpret; we used the remaining question to measure distributive justice. Finally, we asked about use of social media platforms and demographic questions. For demographics, gender and race categories were not mutually exclusive; each respondent could choose multiple options, and many fell into multiple categories (e.g., non-binary and woman).

Recruitment and Demographics

We recruited participants via Prolific, Mechanical Turk, Positly, and word-of-mouth. Workers on MTurk had to have a HIT approval rate of higher than 98%, be located in the U.S., and have had more than 5,000 HITs approved. We compensated Prolific, MTurk, and Positly respondents $3. Word-of-mouth respondents were given the option of receiving a $3 Amazon gift card or donating $3 to charity; most chose the gift card. We removed low quality responses and our final sample had 573 participants. The median duration to complete the survey was 7.5 minutes; the 75% quartile duration was just over 10 minutes, which confirms a median wage of more than $15/hour. We recruited participants in batches to sample diverse demographics; our final sample included people who were: women (45%), non-binary (9%)[2], transgender (15%), custom gender (1%, including Prefer Not to Disclose and Prefer to Self-Describe), Black or African American (21%), Hispanic or Latino (13%), Asian (8%), American Indian or Alaska Native (4%), Middle Eastern (1%), Native Hawaiian or Pacific Islander (1%), custom race/ethnicity (0%, including Prefer Not to Disclose and Prefer to Self-Describe), Liberal (57%), Conservative (18%), and with household income less than $50k (43%). Liberals are overrepresented due to targeted sampling based on other identities. Participants could choose multiple gender, race/ethnicity, and sexual orientation categories.

[2] Some non-binary participants also considered themselves transgender (*N*=30) and some did not (*N*=13).

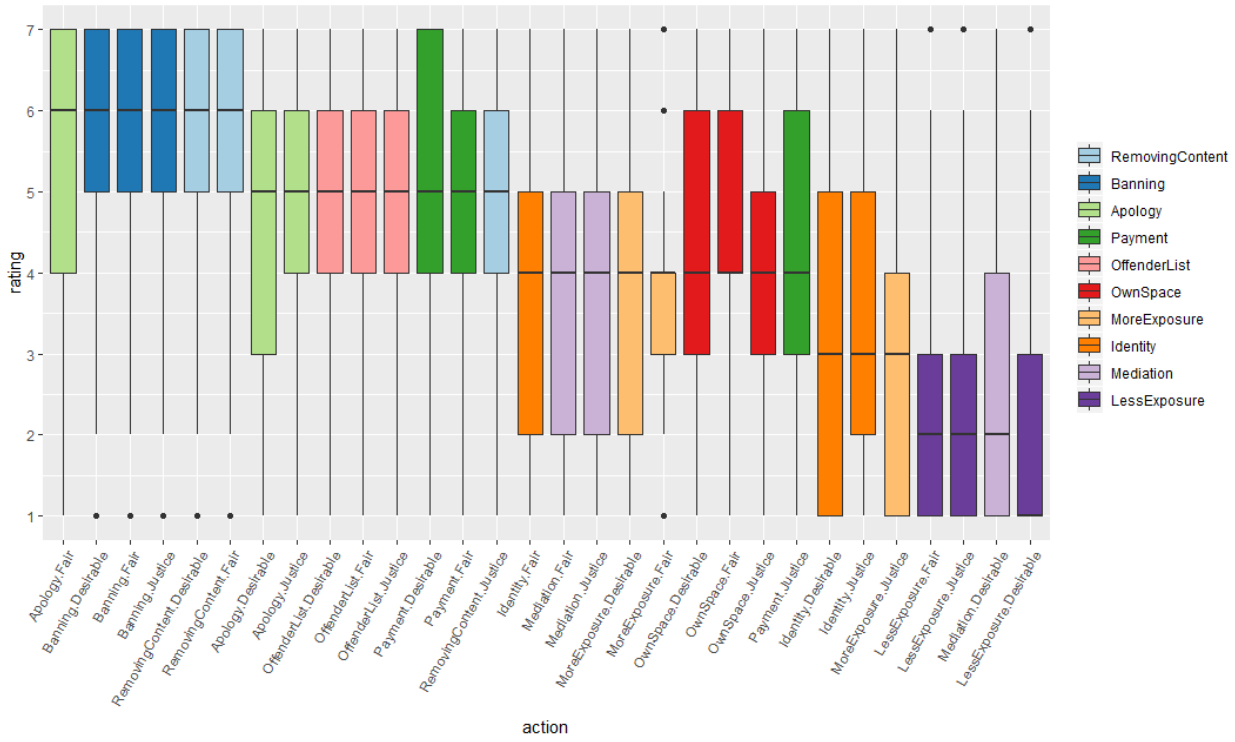
Results

To answer RQ1, we used one-way repeated measures ANOVA with Tukey HSD post-hoc comparisons. Results are shown visually in Figure 1 and multiple comparisons are in supplemental materials. To test H1 and address RQ2-9, we fitted a series of linear regression models modeling preferences as the dependent variable and identity and social media uses/preferences as independent variables. Variance inflation factor (VIF) values were less than three, indicating multicollinearity was not an issue. We used Akaike's Information Criteria (AIC) to exclude independent variables that did not improve the model fit. Models are in Tables 2-5 and communicated via heat map using correlation coefficients in Figure 2. Reference categories are man for gender and white for race/ethnicity.

Preferred Actions

Group means were statistically significantly different from one another [$F(29,17160) = 212.6$, $p<0.0001$]. In general, participants were significantly more likely to favor *banning users* (justice: M=5.26, SD=1.67; fair: M=5.53, SD=1.50; desirable: M=5.46, SD=1.62), *removing content* (justice: M=4.76, SD=1.66; fair: M=5.21, SD=1.54; desirable: M=5.21, SD=1.64), and *apology* (justice: M=4.70, SD=1.82; fair: M=4.93, SD=1.62; desirable: M=4.61, SD=1.82) actions over other actions. They were most opposed to the *less exposure* (justice: M=2.67, SD=1.73; fair: M=2.77, SD=1.78; desirable: M=2.79, SD=1.85) and *more exposure* (justice: M=3.34, SD=1.69; fair: M=3.75, SD=1.56; desirable: M=3.74, SD=1.72) actions.

Post hoc comparisons indicated that justice, fairness, and desirability ratings of an action were not statistically significantly different from one another for the *banning*, *offender list*, *own space, identity, and less exposure* actions. Fairness was rated significantly more highly than justice in four conditions (and higher but not statistically significant in the other six): removing content (justice: M=5.26, SD=1.67; fair: M=5.53, SD=1.50), payment (justice: M=4.15, SD=1.89; fair: M=4.58, SD=1.72), own space (justice: M=4.05, SD=1.79; fair: M=4.37, SD=1.68), and more exposure (justice: M=3.34, SD=1.69; fair: M=3.75, SD=1.56). Other actions revealed differences, for example, the *payment* action was significantly more desirable (M=4.72; SD=1.90) than just (M=4.15; SD=1.89), whereas the *mediation* action was significantly more fair (M=3.78; SD=1.77) than desirable (M=3.52; SD=1.92).



Preferences by Identity

Women were significantly more likely to respond that the *mediation* action would *not* restore justice to them, and that the *more exposure* action was undesirable (see Models 19, 12). That is, on average, women did not desire increased exposure to a large audience.

Transgender participants were significantly more likely to consider the *identity* action fair and desirable (Models 23, 24). However, they were less likely to favor *payment* on all dimensions: justice, fairness, and desirability (Models 28-30). Transgender participants were less likely to find *more exposure* just or desirable but considered the *apology* action less desirable (Models 10, 12, 18).

Non-binary participants were more likely to report that the *own space* action would restore justice to them, and the *removing content* action was fair (Models 13, 2). They were less likely to consider the *less exposure* action fair or desirable (Models 8, 9).

Black participants reported the *more exposure* action as less just relative to white participants (Model 10). American Indian or Alaska Native participants were more likely to report the *identity* action to be fair, but the *payment* action to be unfair and the *banning* action to be undesirable (Models 23, 29, 6). Hispanic or Latino participants were more likely to find the *payment*, *apology*, and *own space* actions to be unfair (Models 29, 17, 14). Asian participants were more likely to favor the *mediation* and *identity* actions as just, fair, and desirable (Models 19-21, 22-24) and the *offender list* and *apology* as desirable (Models 27, 18).

People with a higher household income were significantly more likely to find *payment* to be unfair and *own space* to be undesirable (Models 29, 15). Neither education nor employment predicted any of the actions. Older adults were more likely to disfavor the *identity* actions as not just, unfair, and undesirable (Models 2-24). They also found the *more exposure* and *own space* actions as not just and also undesirable (Models 10, 12, 13, 15).

Finally, liberal participants (as opposed to conservative participants) were significantly more likely to favor the *identity* actions as just, fair, and desirable (Models 22-24). They were more likely to find the *removing content* and *banning* actions to be fair and desirable (Models 2, 3, 5, 6), the *own space* action to be fair, and the *payment* action to be desirable (Models 14, 30).

Preferences by Social Media Use

Instagram users were significantly more likely to consider the *removing content* and *offender list* actions as just, fair, and desirable (Models 1-3, 25-27), the *payment* actions as just and fair (Models 28, 29), and the *banning users* action as fair (Model 5). They were more likely to oppose *less exposure*, reporting it as less fair and less desirable (Models 8, 9). Reddit users were more likely to support the *removing content* action as fair but opposed *payment* as neither just nor desirable (Models 2, 28, 30). They also were more likely to report *less exposure* actions to be unfair and undesirable (Models 8, 9), as well as the *own space* actions as not just and the *offender list* undesirable (Models 13, 27).

Twitter users were more likely to be opposed to the *offender list*, *apology*, and *mediation* actions, reporting each of these as less just, fair, and desirable (Models 25-27, 16-21). Facebook users were more likely to support the *identity* action as just, fair, and desirable (Models 22-24). They also found the *apology* action to be just (Model 16). Snapchat users found the *offender list* action to be undesirable (Model 27). YouTube users supported the *more exposure* action as fair (Model 11).

Preferences based on Prior Experiences with Harassment

Participants who had been harassed themselves were more likely to support the *more exposure* action as just and desirable (Models 10, 12), and those who have been harassed based on their identity were more likely to find the *more exposure* action to be fair and the *less exposure* action to be unjust (Models 11, 7).

Participants who had supported harassment targets were significantly more likely to support the *own space, apology, and mediation* actions as just, fair, and desirable (Models 13-21). They were more likely to report that the *banning* and *identity* actions were just and desirable (Models 4, 6, 22, 24). They also were more likely to report the *more exposure* and *removing content* actions as desirable (Model 12, 3).

Participants who had harassed others were significantly more likely to oppose the *removing content* and *banning* actions across all measures—justice, fairness, and desirability (Models 1-6). They also were more likely to report that the *less exposure* action was both just and fair, while the *payment* action was desirable and educating people about their *identity* was fair (Models 7, 8, 30, 23).

Preferences by Orientations towards Justice

People who feel that social media sites distribute outcomes fairly after harassment (i.e., distributive justice) were significantly more likely to find the *removing content* action to be just, but found both the *payment* and *apology* actions to be unfair (Models 1, 29, 17). People who feel that social media sites treat people fairly after harassment (i.e., procedural justice) were more likely to be supportive in general of most response options (see details in Models 4, 6-8, 10-15, 27, 29).

Research Question/Hypothesis	**Supported?**
RQ1 Are there differences in attitudes towards perceived a) justice, b) fairness, and/or c) desirability of traditional versus alternative actions taken by social media sites?	**Yes** Traditional actions generally viewed as more just, fair, and desirable than alternative actions, with a few exceptions (e.g., *apology, offender list*). See Figure 1, supplemental materials.
H1a[b] Non-dominant [Dominant] social groups will be more favorable towards perceived a) justice, b) fairness, and/or c) desirability of alternative [traditional] actions taken by social media sites, as opposed to traditional [alternative] actions.	**Partially** Transgender participants more likely to support *identity* action. Non-binary participants more likely to support *own space* action. American Indian or Alaskan Native participants more likely to support *identity* action and less likely to support *banning users.* Asian participants more likely to support *identity*, *mediation*, *offender list*, and *apology* actions. See Figure 2, Tables 2-5.
RQ2-6 Is [race/ethnicity [RQ2] / gender [RQ3] / political orientation [RQ4] / socioeconomic class [RQ5] / age [RQ6]] associated with attitudes towards perceived a) justice, b) fairness, and/or c) desirability of actions taken by social media sites?	**Partially** Liberal participants more likely to support *banning users*, *removing content*, *identity*, and *own space*. High SES participants less likely to support *payment* and *own space*. Older participants more likely to support *less exposure* and less likely to support *identity* and *own space*. See Figure 2, Tables 2-5, and the above cell.

RQ7 Is frequency of use associated with attitudes towards perceived a) justice, b) fairness, and/or c) desirability of actions taken by social media sites?	**Partially** Instagram users more likely to support *removing content* and *offender list* but not *less exposure*. Twitter users less likely to support *offender list*, *apology*, and *mediation*. Facebook users more likely to support *identity*. See Figure 2, Tables 2-5.
RQ8 Are prior experiences of harassed or harassing others on social media associated with attitudes towards perceived a) justice, b) fairness, and/or c) desirability of actions taken by social media sites?	**Yes** Participants who had been harassed more likely to support *more exposure*. Participants who had harassed others more likely to support *less exposure* and *identity*, less likely to support *removing content.* See Figure 2, Tables 2-5.
RQ9 Are orientations towards procedural or distributive justice associated with attitudes towards perceived a) justice, b) fairness, and/or c) desirability of actions taken by social media sites?	**Yes** Participants who support distributive justice less likely to support *payment* and *apology*. Participants who supported procedural justice more likely to support most actions, except for *removing conten*t.

Table 6. Summary of results.

Discussion

Apology: Visibility and Reparation

The *apology* action was strongly supported by participants—rated highly as fair, as well as just and desirable. Social media sites requiring apologies from offenders would indicate to targets that the social media site deemed the offense to be in violation of appropriate behavior. However, while apologies evoke acceptance of responsibility and remorse in American discourse, they are notably absent from dispute resolution and legal systems in the U.S. (Wagatsuma & Rosett, 1986). When present in judicial processes, apologies are usually delivered in exchange for automated sentence reductions for guilty pleas (Bibas & Bierschbach, 2004). They also may invoke admission of guilt, thus increasing punishment (Scher & Darley, 1997). Social media sites could apologize to targets which would acknowledge harms they have experienced. They could also require apologies from offenders which might increase support and visibility for targets (which would be preferable for some groups but not others), while also enabling a gradated sanction before banning a user entirely. Though the idea of an apology is not baked into social media governance, volunteer moderators sometimes solicit apologies from offenders, or deliver apologies to targets themselves, indicating this approach's potential (Seering et al, 2019; Matias, 2019).

Open questions remain regarding how apologies should be delivered and whether they need to be genuine. Reparation requires that apologies contain specific linguistic signals, including expressions of responsibility and remorse (Scher & Darley, 1997). However, whether and how apologies restore justice and fairness to targets is likely to vary by identity. Transgender participants rated the apology as not desirable and Hispanic or Latino participants rated it as unfair, perhaps because an ingenuine apology would be overtly harmful and would magnify discriminations those groups experience. Twitter users rated the apology as neither just, fair, nor desirable, perhaps because they felt that an ingenuine apology is not likely to occur on Twitter, and because it could be coopted for further harassment. People who had supported harassment targets were strongly supportive of the apology, which reflects their allyship towards targets.

Public Shaming

Participants rated the online *offender list* action as just, fair, and desirable; however, public shaming has been generally discarded in legal scholarship as subversive to human equality and dignity (Nussbaum, 2009). Shaming labels a person as bad instead of labeling the person's act as bad, thus marking a person with a degraded identity within society (Nussbaum, 2009). Shaming sanctions may

further inflict their greatest weight on marginalized groups, magnifying the penalties on dignity (Nussbaum, 2009). In offline contexts in the U.S., shaming has been used sporadically for low-level crimes (e.g., standing on a street corner with an affixed sign), as well as for more severe crimes—notably, sex offenses, which require listing on a public registry. However, public registries are aimed at community protection rather than punishment, and seek to strike a balance between protecting basic civil liberties as guaranteed by the Constitution and protecting the public from harm.

Participants' desire for an offender list may reflect a visceral, even primal, desire to punish offenders in the absence of visible or effective penal systems on social media (Blackwell, Chen, Schoenebeck, & Lampe, 2018). Instagram and Twitter users were more likely to rate the offender list as just, fair, and desirable—indicating a strong orientation towards public approaches to norm enforcement on those platforms. Snapchat users rated offender lists as undesirable, which reflects Snapchat's typically small, tight-knit communities (Bayer, Ellison, Schoenebeck, & Falk, 2015). Klonick (2015) raises three overarching concerns that public shaming on the Internet: is not a calibrated or measured form of punishment, has questionable accuracy in terms of who or what it punishes, and results in an over-determined punishment with indeterminate social meaning. In other words, low cost, anonymous, instant, and easy access to the Internet has eviscerated whatever "natural" limits there were to public shaming and has served to amplify its effects (Klonick, 2015). However, online shaming can be used effectively if it shames the violation rather than the norm violator (Klonick, 2015). Social media sites could sanction behavior by making online harassment cases publicly visible, but without identifying the offender.

The Limitations of One-Size-Fits-All Approaches

Our results reveal how a one-size-fits-all approach to online harassment may fail to support some users while privileging others. For example, while *banning users* was popular overall, American Indian or Alaska Native participants considered banning users undesirable. This may reflect this group's cultural preference for restorative rather than retributive justice, their historical experiences of being forcibly removed from their own land (Brave Heart & DeBruyn, 1998), or their recent history of Facebook account bans due to names misaligned with the site's "real name" policies (Haimson & Hoffmann, 2016). The site action of educating other users about an individual's *identity* was favorable to some marginalized groups: participants who were transgender, American Indian or Alaska Native, and Asian. These are groups whose identities are frequently misunderstood, or even feared, in mainstream society, and who may tire of educating others about who they are and how they wish to be treated (Brave Heart & DeBruyn, 1998; James et al., 2016). Transgender, American Indian or Alaska Native, and Hispanic or Latino participants deemed *payment* as neither just, fair, nor desirable, despite these groups all facing substantial economic disparities in the U.S. (Brave Heart & DeBruyn, 1998; James et al., 2016; Patten, 2016). These groups may perceive compensation as a band-aid that overlooks and undervalues, rather than addresses, the racial, economic, or social injustices they experience. Payment may also be, again, at odds with American Indian or Alaska Native people's community-oriented rather than compensatory approaches to justice (Brave Heart & DeBruyn, 1998; Melton, 1995).

Participants who had been harassed previously were favorable to *more exposure* as just, fair, and desirable. This aligns with Citron's (2014) suggestion that harassment targets may benefit from more exposure on social media sites, such as receiving discounted advertising rates to clear their reputation, or dispute negative things said about them by harassers. However, some groups—transgender people, Black people, and women—found these solutions less just and desirable. It could be that if exposed to a nonconsensual spotlight, some may wish to remove themselves from the public eye rather than gaining a larger audience. In the case of transgender people, widespread disclosure of their trans identity may render them especially vulnerable to violence and discrimination in the physical world. However, non-binary participants responded differently—they felt that *less exposure* would be unfair and undesirable. While there is much overlap between transgender and non-binary participants, it could be that non-binary people often assert their

identities (e.g., appearance, pronouns) in visibly non-binary ways and less exposure would limit those assertions.

Our results lay the groundwork for how, and why, social media sites should consider identities and social groups when determining online harassment processes and policies. Indeed, in her reflections on the U.S. justice system's treatment of Black women, Crenshaw (1991) observed that removing differences between people overlooks their unique identities and experiences. Further, rather than transcending such differences (as social media sites' mantra of neutrality might claim), one-size-fits-all approaches instead flatten intragroup differences and magnify structural inequities in experiences of justice (Crenshaw, 1991). Our work only focused on U.S.-perspectives; it is likely that a monolithic approach to governance further magnifies inequities when applied in global, cross-cultural contexts.

Our arguments are thus two-fold: that **justice should be the principled foundation on which social media governance decisions are made, and that justice can be integrated into the design of social media systems**. For example, in the U.S., while apologies are typically absent from judicial systems (Wagatsuma & Rosett, 1986), site apologies to targets of harassment could be more closely integrated into the content moderation process via a combination of automated and human processes. Other approaches, like payment, align with existing social norms and technological infrastructures on sites like Instagram and could similarly be integrated into governance processes. This work was motivated by the concern that criminal justice approaches in the U.S. are limited in their ability to reform offenders, and exacerbate inequities based on identity. We aim to inspire reflection and action into the merits of some criminal justice approaches, and the possibilities opened up by alternative justice theories, to support targets online.

Conclusion and Future Work

This study presents a broad range of social media site responses to online harassment and considers their potential for supporting targets. We put forth a theoretical argument for the limitations of the criminal justice models for supporting targets, and consider alternative approaches that recognize systematic and structural power imbalances. Our study focused only on U.S. perspectives which represents a narrow slice of global social media use; future work in other regions of the world could bring alternative, promising approaches to justice and social media governance. Future work could also examine how participants across cultures interpret the concepts of justice and fairness. An additional limitation is that the current study did not test efficacy of proposed solutions. In some cases, like mediation, people may not like the idea in theory but may find it restorative in practice. Our study intentionally focused on the wellbeing of online harassment targets, because they have been overlooked to date. As a result, we did not measure attitudes towards what may be best for the offender, or for the community. If we had measured morality or dignity for all involved parties, we might observe reduced support for some approaches, like offender lists, which can indiscriminately penalize offenders.

Social media sites want to present themselves as neutral arbiters of online content (Gillespie, 2010); however, such arbitration procedures can differentially impact social media users based on their individual identities and experiences. Our results indicate opportunities for developing alternative theories and approaches to supporting targets with more just, fair, and desirable responses to online harassment.

References

Ahmed, S. (2012). *On Being Included: Racism and Diversity in Institutional Life*. Duke University Press.

Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning. Retrieved August

26, 2019, from https://fairmlbook.org/

Bayer, J., Ellison, N. B., Schoenebeck, S. Y., & Falk, E. B. (2015). Sharing the small moments: Ephemeral social interaction on Snapchat. *Information, Communication & Society*, *19*(7), 1–22. https://doi.org/10.1080/1369118X.2015.1084349

Bell, D. (2008). *And We Are Not Saved: The Elusive Quest For Racial Justice*. Basic Books.

Bibas, S., & Bierschbach, R. A. (2004). Integrating Remorse and Apology into Criminal Procedure Essay. *Yale Law Journal*, (1), 85–148.

Blackwell, L., Chen, T., Schoenebeck, S., & Lampe, C. (2018). When Online Harassment is Perceived to be Justified. *International AAA Conference on Web and Social Media (ICWSM 2018)*. AAAI Press.

Blackwell, L., Dimond, J., Schoenebeck, S., & Lampe, C. (2017). Classification and its Consequences for Online Harassment: Design Insights from HeartMob. *Proc. ACM Hum.-Comput. Interact.*, *1*(2), 19 pp.

Bobo, L. D., & Thompson, V. (2006). Unfair by Design: The War on Drugs, Race, and the Legitimacy of the Criminal Justice System. *Social Research: An International Quarterly*, *73*(2), 445–472.

Bradford, B., Grisel, F., Meares, T. L., Owens, E., Pineda, B. L., Shapiro, J., … Peterman, D. E. (2019). *Report Of The Facebook Data Transparency Advisory Group*. Yale Law School: Yale Justice Collaboratory.

Braithwaite, J. (1999). Restorative Justice: Assessing Optimistic and Pessimistic Accounts. *Crime and Justice*, *25*, 1–127. https://doi.org/10.1086/449287

Brave Heart, M. Y., & DeBruyn, L. M. (1998). The American Indian Holocaust: Healing historical unresolved grief. *American Indian and Alaska Native Mental Health Research: Journal of the National Center*, *8*(2), 56–78.

Cai, J., & Wohn, D. Y. (2019). What are Effective Strategies of Handling Harassment on Twitch? Users' Perspectives. *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*, 166–170. https://doi.org/10.1145/3311957.3359478

Caplan, R. (2018). *Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches*. Data & Society.

Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. Crossmod: A Cross-Community Learning-based System to Assist Reddit Moderators. Proc. ACM Hum.-Comput. Interact. 3, CSCW, Article 174 (November 2019), 30 pages. https://doi.org/10.1145/3359276

Citron, D. K. (2014). *Hate Crimes in Cyberspace*. Harvard University Press.

Clayton, S., & Opotow, S. (2003). Justice and Identity: Changing Perspectives on What Is Fair. *Personality and Social Psychology Review*, *7*(4), 298–310. https://doi.org/10.1207/S15327957PSPR0704_03

Coates TN (2014) The case for reparations. *The Atlantic* 313(5): 54–71.

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

Crenshaw, K. (1991). Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. *Stanford Law Review*, *43*(6), 1241–1299. https://doi.org/10.2307/1229039

D'Ignazio, C., & Klein, L. (2020). *Data Feminism.* MIT Press.

Duggan, M. (2017, July 11). Online Harassment 2017. Retrieved September 17, 2018 from http://www.pewinternet.org/2017/07/11/online-harassment-2017/

Fallon, Richard H., & Weiler, P. C. (1984). Firefighters v. Stotts: Conflicting Models of Racial Justice. *The Supreme Court Review*, *1984*, 1–68. https://doi.org/10.1086/scr.1984.3536937

Gillespie, T. (2010). The politics of 'platforms.' *New Media & Society*, *12*(3), 347–364. https://doi.org/10.1177/1461444809342738

Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press.

Haimson, O. L., & Hoffmann, A. L. (2016). Constructing and enforcing "authentic" identity online: Facebook, real names, and non-normative identities. *First Monday*, *21*(6). https://doi.org/10.5210/fm.v21i6.6791

Hoffmann, A. L. (2019). Where fairness fails: Data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society*, *22*(7), 900–915. https://doi.org/10.1080/1369118X.2019.1573912

Hoffmann, A. L., & Jonas, A. (2016). Recasting Justice for Internet and Online Industry Research Ethics. In M. Zimmer & K. Kinder-Kuranda (Eds.), *Internet Research Ethics for the Social Age: New Cases and Challenges*. Retrieved from https://papers.ssrn.com/abstract=2836690

Jackson, T. F. (2013). *From Civil Rights to Human Rights: Martin Luther King, Jr., and the Struggle for Economic Justice*. University of Pennsylvania Press.

James, S., Herman, J., Rankin, S., Keisling, M., Mottet, L., & Anafi, M. (2016). *The Report of the 2015 U.S. Transgender Survey*. Retrieved from https://ncvc.dspacedirect.org/handle/20.500.11990/1299

Jhaver, S., Ghoshal, S., Bruckman, A., & Gilbert, E. (2018). Online Harassment and Content Moderation: The Case of Blocklists. *ACM Trans. Comput.-Hum. Interact.*, *25*(2), 12:1–12:33. https://doi.org/10.1145/3185593

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. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 192 (November 2019), 33 pages. DOI: https://doi.org/10.1145/3359294

Klonick, K. (2015). Re-Shaming the Debate: Social Norms, Shame, and Regulation in an Internet Age. *Maryland Law Review*, *75*, 1029.

Laloggia, J., & Inquiries. (2019). *U.S. public has little confidence in social media companies to determine offensive content*. Retrieved from https://www.pewresearch.org/fact-tank/2019/07/11/u-s-public-has-little-confidence-in-social-media-companies-to-determine-offensive-content/

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

Matias, J. N. (2019). The Civic Labor of Volunteer Moderators Online. Social Media + Society. https://doi.org/10.1177/2056305119836778

Matias, J. N., Johnson, A., Boesel, W. E., Keegan, B., Friedman, J., & DeTar, C. (2015). *Reporting, Reviewing, and Responding to Harassment on Twitter* (SSRN Scholarly Paper No. ID 2602018). https://papers.ssrn.com/abstract=2602018

Melton, A. P. (1995). Indigenous Justice Systems and Tribal Society Indian Tribal Courts and Justice: A Symposium. *Judicature*, *79*, 126–133.

Munger, K. (2017). Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment. *Political Behavior*, *39*(3), 629–649. https://doi.org/10.1007/s11109-016-9373-5

Nelson A (2016) The Social Life of DNA: Race, Reparations, and Reconciliation After the Genome. Boston, MA: Beacon Press.

Nussbaum, M. C. (2009). Hiding from Humanity: Disgust, Shame, and the Law. Princeton University Press.

Pater, J. A., Kim, M. K., Mynatt, E. D., & Fiesler, C. (2016). Characterizations of Online Harassment: Comparing Policies Across Social Media Platforms. *Proceedings of the 19th International Conference on Supporting Group Work*, 369–374. https://doi.org/10.1145/2957276.2957297

Patten, E. (2016). *Racial, gender wage gaps persist in U.S. despite some progress*. Retrieved from https://www.pewresearch.org/fact-tank/2016/07/01/racial-gender-wage-gaps-persist-in-u-s-despite-some-progress/

Postigo H (2016) The socio-technical architecture of digital labor: Converting play into YouTube money. *New Media & Society* 18(2): 332–349.

Rawls, J. (2009). *A Theory of Justice*. Harvard University Press.

Roberts, S. (2019). *Behind the Screen: Content Moderation in the Shadows of Social Media*. Retrieved from https://www.amazon.com/Behind-Screen-Content-Moderation-Shadows/dp/0300235887

Scher, S. J., & Darley, J. M. (1997). How Effective Are the Things People Say to Apologize? Effects of the Realization of the Apology Speech Act. *Journal of Psycholinguistic Research*, *26*(1),

127–140. https://doi.org/10.1023/A:1025068306386

Seering, J., Wang, T., Yoon, J., & Kaufman, G. (2019). Moderator engagement and community development in the age of algorithms. New Media & Society, 21(7), 1417–1443. https://doi.org/10.1177/1461444818821316

Sunshine, J., & Tyler, T. R. (2003). The Role of Procedural Justice and Legitimacy in Shaping Public Support for Policing. *Law & Society Review*, *37*(3), 513–548. https://doi.org/10.1111/1540-5893.3703002

The Combahee River Collective Statement. (n.d.). Retrieved from http://circuitous.org/scraps/combahee.html

Tyler, T. R. (1994). Psychological models of the justice motive: Antecedents of distributive and procedural justice. *Journal of Personality and Social Psychology*, *67*(5), 850–863. https://doi.org/10.1037/0022-3514.67.5.850

Wagatsuma H and Rosett A (1986) The implications of apology: Law and culture in Japan and the United States. *Law & Sociology Review* 20: 461.

Wenzel, M., Okimoto, T. G., Feather, N. T., & Platow, M. J. (2008). Retributive and Restorative Justice. *Law and Human Behavior*, *32*(5), 375–389. https://doi.org/10.1007/s10979-007-9116-6

Donghee Yvette Wohn. 2019. Volunteer Moderators in Twitch Micro Communities: How They Get Involved, the Roles They Play, and the Emotional Labor They Experience. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19). ACM, New York, NY, USA, Paper 160, 13 pages. DOI: https://doi.org/10.1145/3290605.3300390

Zehr, H. (2015). *The Little Book of Restorative Justice: Revised and Updated*. Simon and Schuster.