Youth Trust in Social Media Companies and Expectations of Justice: Accountability and Repair after Online Harassment

SARITA SCHOENEBECK, University of Michigan, USA
CAROL F. SCOTT, University of Michigan, USA
EMMA HURLEY, University of Michigan, USA
TAMMY CHANG, University of Michigan, USA
ELLEN SELKIE, University of Michigan, USA

Social media platforms aspire to deliver fair resolutions after online harassment. Platforms rely on sanctions like removing content or banning users but these punitive responses provide little opportunity for justice or reparation for targets of harassment. This may be especially important for youth, who experience pervasive harassment which can have uniquely harmful effects on their wellbeing. We conducted a text-message based survey with 832 U.S. adolescents and young adults, ages 14-24, to explore their attitudes towards social media companies' responses to online harassment. We find that youth are twice as likely (41% versus 20%) not to trust social media companies' ability to achieve a fair resolution as they are to trust them. Nearly two-thirds (62%) of youth expressed a preference for an apology from the offender after online harassment, and they were twice as likely to prefer a private apology to a public one (29% versus 14%). Preferences also vary by identity, revealing how a one-size-fits-all approach can harm some youth while benefitting others. We reflect on the opportunities and risks associated with institutional trust and restorative justice for supporting youth who experience online harassment.

CCS Concepts: • **Human-Centered Computing** → Collaborative and social computing; **Human-Centered Computing** → Social media

KEYWORDS: Content moderation, reparative justice, restorative justice, criminal justice, punishment, fairness, youth, trust, online harassment

ACM Reference format:

Sarita Schoenebeck, Carol F. Scott, Emma Hurley, Tammy Chang, Ellen Selkie. 2021. Youth Trust in Social Media Companies and Expectations of Justice: Accountability and Repair after Online Harassment. In *Proceedings of the ACM on Human-Computer Interaction,* Vol. 5, CSCWXX, Article 1 (2021), 18 pages, https://doi.org/10.1145/3449076

1 INTRODUCTION

Fairness is a central principle in the U.S. justice system [73]. Fairness, broadly speaking, refers to the absence of bias in procedures and processes. Fairness has been enacted in criminal justice systems through an approach called procedural justice, which seeks to increase transparency, and thereby compliance, of decision-making processes [73]. Social media companies have similarly aspired to embrace fairness in how they respond to online harassment on their platforms [5, 18,

2573-0142/2021/April - 2 $15.00

https://doi.org/10.1145/3449076

19]. These aspirations towards fairness are intended to create equal, non-discriminatory experiences and to engender trust among social media users [32, 35, 65, 70]. However, Wanda McCaslin writes in her critique of Western law, "One colonizer response is to appeal to values of equality and fairness…. Colonizer fairness means 'imposing one law for all.' But whose law is the one to be imposed? Who is favored and who is constrained by the 'one law'? [43]. Though responses to online harassment are advancing, most social media companies have struggled to effectively fair or equitable procedures, or to deliver fair or equitable outcomes, and many experts believe the problems are getting worse [54].

Online harassment refers to a wide range of behaviors that seek to threaten, harm, or disparage another person or group. Online harassment correlates with a variety of negative outcomes, including poor health, relationships, and job security as well as the degradation of civil discourse, justice, and general welfare [52, 54, 61]. Most social media platforms maintain community guidelines that dictate what kinds of behaviors are or are not appropriate on the site. To determine whether content on their platforms violates community guidelines, social media companies rely on a process called content moderation, which assesses posts using both automated and human methods [21, 60, 66]. If content is found to violate community guidelines, most platforms respond by removing content or sanctioning offenders, typically via warnings or bans [65]. However, these models largely write the targets of online harassment out of the justice-seeking process. On social media platforms, if a user reports harassment, they may receive a response indicating whether an action was taken, but are given little opportunity for visibility or reparation [65].

This research considers how justice theories can inform how social media companies and communities respond to online offenses among youth [46, 77]. It also critically reflects on existing orientations towards criminal justice with its commitments to punitive sanctions, and examines emerging alternative approaches like restorative justice. Restorative justice and transformative justice activists Mariame Kaba and Mia Mingus have advocated for abolishment of carceral punishment systems and towards systems of accountability [36, 47, 79]. Mariame Kaba poses the question: "what would be a just system for adjudicating and evaluating harm?" She continues: "It's a question that invites people in, that invites people to offer their ideas… rather than accepting as permanent and always necessary the current oppressive institutions that we have" [36]. Scholar Mariam Asad has similarly proposed community intervention approaches to repair harm as alternatives to contemporary punitive criminal justice systems [1].

Restorative justice is a countermovement to criminal justice and works by bringing offenders, targets, community or family members, and mediators together to acknowledge and remediate harms [6, 10, 77]. A growing collective of scholars in the U.S.—where the current study took place—are examining intersections between alternative justice approaches and technology, including Asad [1], Amy Hasinoff, Anna Gibson, and Niloufar Salehi [63], Lindsay Blackwell and colleagues [49], and in the first author's prior work with Oliver Haimson and Lisa Nakamura [65]. Restorative justice approaches have been particularly effective with some populations of youth because their psychological development matures substantially through childhood and adolescence, allowing for learning and rehabilitation rather than solely punishment [51]. Additionally, restorative justice approaches recognize racial and social injustices that lead to over-incarceration of young people. In New Zealand, for example, restorative justice was the foundation for a 1989 act between Maori people and the New Zealand Parliament which was designed to care for Indigenous children rather than moving them into prison pipelines. This research takes a youth advocacy lens, and considers how social media platforms can better support youth, who might be especially susceptible to risks and harms associated with harassment [42, 53].

We conducted a text message based free-response survey with 832 young adults, ages 14-24, to learn about their trust in social media companies' responses to online harassment, and their preferences for fair resolutions. As part of the advocacy lens, our study sought a demographic of youth who are diverse in terms of race and socioeconomic status, because those groups may be

more likely to distrust institutions and may be less likely to be supported when they experience harm. Our research team draws on expertise from computing, social work, family medicine, and adolescent health to center youth wellbeing in their own experiences. This work contributes to scholarship in Internet research by advancing alternative justice theories in the context of online harassment, and by prioritizing youth's own voices in how platforms should respond to online harassment. It also contributes new insights for healthcare practitioners, social workers, and caregivers to understand youths' online experiences of harm and restoration.

2 PRIOR WORK

Trust is a complex concept with an expansive associated body of literature. We focus our review here on youth trust in institutions as it relates to how they might trust social media companies. Fairness also spans multiple disciplines and has numerous definitions and operalizations. We scope our review to a summary of fairness as it relates to content moderation and principles of justice.

2.1 Youth Trust in Institutions

Friedman et al.'s 2000 article on trust online opens with: "Trust matters. It allows us to reveal vulnerable parts of ourselves to others and to know others intimately in return. A climate of trust eases cooperation among people and fosters reciprocal caretaking. The resources—physical, emotional, economic—that would otherwise be consumed guarding against harm can be directed toward more constructive ends" [20]. They describe trust as something that is present when there is an opportunity for another person to cause harm, but one feels confident the other person will not take that opportunity [20]. Trust, and the reciprocation of trust with trustworthiness, has benefits for economic growth, social cohesion, physical health, and subjective wellbeing [25, 26]. Youth who trust in others and trust in government are more likely to participate in community service, voting, and political volunteerism. However, increased trust can sometimes increase risk—among older adults, increased trust may increase incidences of financial exploitation, health care fraud, and digital deception [34].

Younger adults tend to have lower trust than older adults. Almost half of young adults (46%) are "low trusters" —people who are more likely to see others as selfish, exploitative and untrustworthy, rather than helpful, fair and trustworthy—compared with 19% ages 65+ [82]. Differences in trust may be a cohort or generational phenomenon, though trust may also generally increase as people age [34]. Socioemotional selectivity theory suggests that as time horizons grow smaller, people become more selective in how they expand time and resources [8]. Younger adults may exhibit an expansive view of the future that focuses on gaining information and knowledge, whereas older adults focus on emotion and wellbeing and trusting others [39]. Younger adults may also value fairness more, and react more strongly to unfair treatment [2].

Young adults have complicated relationships with perceptions of fairness and of trust in others and in institutions. In general, young adults are less trusting of military, religious, police, and business leaders than older adults (though they are more trusting of scientists, journalists, and college professors) [82]. People of color, people with lower incomes, and people with less education are also less likely to trust others. Noticeably, white people are twice as likely (27%) to be high trusters than people of color (13% Black; 12% Hispanic) [55]. Youth trust in police presents a case study to explain some of these differences. Young people's trust in police begins forming at a young age and police-youth interactions have emerged as a problematic relationship. Surveillance of young people of color, neighborhood policing, and discrimination have all dismantled trust between youth and police [31]. Authoritarian approaches with youth have also lessened youth trust in leaders and institutions (e.g., in schooling systems). However, young people's attitudes towards police legitimacy are positively associated with use of procedural justice. Procedural justice theory demonstrates that people are more likely to find decisions fair,

and to thus comply with them, if they believe the authority making the decision is legitimate [73]. Transparency in the decision-making process also increases the perception of fairness and thus, compliance [69]. Some evidence suggests that procedural justice carries more weight with youth than with adults [48], which may be because they value fairness more [2]. One study indicates that procedural justice approaches may also reduce the likelihood of youth reoffending, at least within a three-month window. In that study, the effect was gone after six months [50], perhaps because youth may need a "booster" that reinforces intervention messaging.

There are fewer studies of young adults' trust in technology companies, though it's possible their levels of trust mirror adults'. In general, Americans tend to have complicated relationships with technology companies. Their positive perceptions of technology companies decreased from 2015 to 2019 [16]. A majority of Americans (55%) said tech companies have too much power and influence, while 72% said it was likely that social media platforms intentionally censor political viewpoints they find objectionable [67]. The public generally believes social media companies have a responsibility to remove offensive content from their platforms, but have little confidence in the companies when it comes to determining what's offensive [38]. However, one poll run by the technology journalism site, The Verge, in 2020 found that Americans believe that Google, Amazon, Apple, Microsoft, Netflix, and YouTube have an overall positive effect on society and that some Americans trust Microsoft (75%), Amazon (73%), and Facebook (41%) to safeguard their personal information [71]. They are more likely to think that Twitter, Slack, Instagram, and Facebook have an overall negative effect on society [71]. Some of the differences in self-reported beliefs may be explained by methodological differences or biases in study design. It may also reflect a version of the privacy paradox, where people express dissatisfaction with technology practices, but still choose to use those technologies for a variety of reasons [3].

Given the importance of youth trust for developing approaches to combat online harassment, our first research question is: *If youth experience bullying or harassment, which social media companies do they trust to support them, and why?*

2.2 Fairness and Justice in Online Harassment

Social media companies struggle to combat online harassment [17]. Online harassment can be difficult to reliably detect, and even if detected, is then difficult to deter [9]. Companies rely on content moderation processes, which are a set of approaches and practices used to remove content that violates community guidelines [21, 60]. Content moderation combines automation with human labor. Automated approaches use machine learning and natural language processing techniques to generate models and classifiers that detect harmful content (e.g., child pornography, hate speech) [9, 33, 81]. Though these approaches continue to improve, they are subject to false positives and true negatives, with some harmful content evading detection while other innocuous content results in a sanction [33]. Additionally, automatic detection efforts are relatively easy to bypass through language modification [33]. Human labor involves workers who process reported content and make decisions about whether it violates their site's community guidelines [21, 60]. Workers are expected to process reports as quickly as possible, and with little context beyond the offending content itself, making it difficult for them to consistently and thoughtfully make decisions. Workers also tend to be undercompensated, especially in light of the traumatizing content they are asked to review as part of their routine work environments [60]. Content moderation work is also taken on by community administrators and moderators, who are unpaid and have to balance complex interpersonal relationships and emotional work [80]. For content moderation workers in particular, much of their work has been behind the scenes and has only become part of mainstream conversations as researchers have exposed the working conditions of content moderators [21, 60].

The combination of imperfect automation and undercompensated human labor has resulted in a complicated environment in which people often do not know why their content was removed, or why someone else's abusive content was not removed [35, 70]. For young people, this uncertainty

can be especially harmful. Social media is a primary form of social interaction and feeling unsafe or being isolated from their social lifelines can be detrimental to their health and wellbeing.

Our prior work has critiqued platforms' focus on criminal justice models of content moderation (e.g., banning users and removing content) and argued for drawing from alternative justice theories, such as social, racial, or restorative justice approaches [65]. Approaches like compensation or apologies may be restorative to targets of online harassment, either in combination with, or in place of, more traditional punitive responses [65]. Restorative justice, in particular, emphasizes two major components—accountability and reparation—which require that offenders are accountable for their actions and that targets should feel that harms have been repaired. Resolution and respect are important outcomes of the restorative justice process, though they may not always be attainable [6, 10]. Restorative justice among youth has become embedded in many Indigenous communities in Australia, New Zealand, Canada, and the US [23, 30, 46, 62]. However, restorative justice principles are relatively new to social media companies. More generally, little is known about youth preferences towards criminal versus restorative justice principles for responding to online harassment. Thus, our second research question is: *If youth experience bullying or harassment, what would feel like a fair resolution to them?*

We use the language of fairness in our study design because it is familiar and accessible to youth, though we critically reflect on the conceptual and algorithmic assumptions perpetuated by the concept of "fairness" in our discussion. We conducted a text-messaged based survey to answer our research questions. Because prior work on social media companies' responses to online harassment has been largely conducted with adult populations, we chose to conduct an exploratory study rather than to develop hypotheses drawn from adult populations which may not represent youth accurately. In the Discussion, we propose opportunities for follow-up studies and hypothesis-testing based on our results.

3 METHODS

3.1 Survey Design

Participants were part of a national cohort of youth participants, ages 14-24, from the MyVoice project [15]. Participants in the cohort were recruited via targeted Facebook and Instagram ads (sample ad: "Earn $1 a week for texting us what you really think!"). Demographic data was collected from participants when they enroll in the cohort and stored for later analysis. Participants' demographic data is linked to their survey responses via a participant identifier. Participants received a one-time incentive of $5.00 (US dollars) for completing the online demographic questions. Textizen, a web-based platform, was used to send the weekly set of questions to participants. For each weekly set of questions, participants received $1.00 each for completing the entire set of questions, which usually consists of 3-5 items.

While the cohort is not a nationally representative sample, recruitment of participants was designed to match national benchmarks based on weighted samples from the 2016 American Community Survey, including age, gender, race/ethnicity, education, family income, and region of the country. The MyVoice project was approved by the Institutional Review Board. Consent was obtained from participants age 18 years and older; parental consent is waived for minor participants to enable equitable recruitment of low-income and at-risk adolescents.

We developed a 5-item text-message survey. The items were designed to be delivered and responded to via text message. The cohort was administered a survey via text message because it is a medium that most youth are familiar with and use actively. While other demographics—i.e. adults older than our sample—might be unlikely to respond to text message questions, this approach is aligned with youth's existing practices. This approach also allowed them to respond in a place that was likely to be safe and comfortable to them, given the relative privacy of mobile phones compared to other mediums like a shared home computer. While text messaging allowed them to express their voices in their own words and own medium, it constrained the number and

nature of questions we could ask. For example, because the survey was only five questions, we did not ask about social media use or propensity to trust. Further, because the national cohort is focused on youth voices and experiences across a variety of topics associated with youth advocacy, social media use was not collected as part of the larger study goals.

Drawing from prior literature and motivated by our research goals, the research team first developed a longer bank of questions. We discussed and revised the wording of the questions and narrowed down to 5 questions total. We then pilot tested the questions with collaborators who have administered prior surveys to this cohort. Finally, we pilot tested the survey with 10 respondents from the cohort and asked them to reflect on the question design. The final design consisted of 5 questions, asked in the same order for all participants. The content analysis and statistical analysis in this paper focuses on two of those questions:

1. If someone bullies or harasses you on social media, which of the following would feel like a fair resolution: private apology via dm, public apology, deleted posts, banning them, support from friends, something else?
2. If someone experiences bullying or harassment, which social media companies do you trust to achieve a fair resolution for you? Why?

We chose to suggest examples in question 1 because our pilot testing suggested that most people, including youth, would have a hard time answering the question without having an idea of what various kinds of resolutions could be. We generated suggestions that reflected criminal justice and reparative justice approaches. We used "dm" to refer to direct message, which is language that is likely to be familiar with most youth. We discussed the use of "bullying" versus "harassment" extensively among the research team. Cyberbullying and bullying are often used in reference to children and adolescents while harassment is often used in reference to older adults—we chose to include both to account for the age range of our sample. We preregistered our study plan and research questions on the Open Science Framework before data collection began.[1]

3.2 Participant Demographics

In August 2019, 1283 youth were sent the survey via text message and 843 responded to the survey. We removed 11 participants from analysis where responses did not make sense or indicated the participant was not attentive. Some participants responded to some questions but skipped other ones; we retained their data in our analysis. The final sample consisted of 832 participants.

The median age of participants was 18 and ranged from 14-24; the mean age was 18.8. Participant gender was 54.4% female, 36.7% male, and 3.2% nonbinary; 4.4% were transgender. Participants were predominantly white (72%, compared with 76% in the US from the 2019 US census [76]), and then Asian (16%, compared with 6% in the US), Black (13%, which reflects the population of the US), Hispanic (12%, compared to 18%), and American Indian (3%, compared with 0.2%). Participation was overrepresented by the Midwest (41%, compared with 21% from the 2019 US census), and underrepresented from the South (24%, compared with 38%). The West and Northeast were slightly underrepresented as well (16% compared with 24%, and 12% compared with 17%, respectively) [74].

Socioeconomic status was measured via participation in the federal Free and Reduced Lunch Program (FRL). Children qualify for FRL at public schools if their household income is below 130% of the poverty level and for a reduced price if it is between 130-185% of the poverty level. Roughly 37% of participants currently or had recently qualified for Free and Reduced Lunch. Most participants (72%) were living at home with a parent/guardian. Their parents/guardians had mostly (61%) received a Bachelor's degree or higher while 34% had not.

[1] https://osf.io/mbruw/?view_only=275b5d23483c491894ed7f4b46a2afff

3.3 Data Analysis

We combined text message responses with demographic data for analysis. Data was linked via a participant identifier and identifiable data was removed. Our data analysis followed an inductive process [72]. The research team read through the data in multiple passes to clean the data and develop the codebook. First, two members of the research team read through all of the raw data to understand participant responses and begin to identify codes and themes. They iteratively achieved consensus through review of data, codes, and emerging themes. Because of the specific nature of each question, we chose to categorize codes by survey question, such that each question had an individual set of codes, though some codes were repeated across questions. The two researchers triple-checked derived codes and asked clarifying questions with the other authors and the larger MyVoice research team.

We used the web-based Reliability Calculator [57] for Ordinal, Interval, and Ratio data to calculate interrater reliability (IRR) using Krippendorff's alpha. Krippendorff (2004) recommends $\alpha \geq .800$ or higher and where tentative conclusions are acceptable, $\alpha \geq .667$ (between $\alpha = .667$ and .800) [44]. For each question, we extracted 50 random responses and two coauthors independently coded each question. For RQ 1, we had three sets of codes, with corresponding Krippendorff's alpha of 0.931, 0.949, and 1. For RQ 2, Krippendorff's alpha was 0.838 and 0.632. We discussed and reconciled disagreements, which is a principle goal during the interreliability process [44] then revised and finalized the codebook. We extract another 50 random responses, not overlapping with the first 50, and two coauthors independently coded again. For RQ 2, Krippendorff's alpha was 0.924 and 1. The final codebook contained items like "Apology – Sincere", "Ban", and "Resilience" with definitions and examples of each. One researcher then coded the remainder of the data and coauthors conducted supplementary coding and checking. Our quantitative analysis of coded data uses descriptive statistics and we ran regressions using the glm() function in R.

Our qualitative data draws on participant quotations to illustrate themes. In most cases, we use exact quotes from participants. As a result, there are some typos or grammatical inconsistencies in the quotes.

4 RESULTS

4.1 Youth Trust in Social Media Companies

Youth are twice as likely to say that they *do not* trust social media companies to achieve fair resolutions after bullying or harassment than that they *do* trust companies. In total, 41% reported that they did not trust companies, whereas 20% reported that they trusted companies. Participants gave a range of explanations for why they did not trust companies: some participants thought social media companies didn't "*do much besides giving a slap on the wrist to people who harass others*" and that individuals would be overlooked as just one person on the site. Others felt that giving better options to report harassment that were "*not just a predetermined option*" might help prevent harassment. One participant explained: "*None, to be quite honest. They don't have any way of reaching out beyond your initial request, and you can report it as a mistake if they refuse but ultimately nothing usually happens*."

Some participants (14%) felt it was not social media companies' responsibility to ensure a fair resolution, saying, for example "*None of them, this is not their responsibility in my opinion.*" Another participant explained in more detail: "*None, bullying is their [users'] responsibility. They [social media companies] are the same as a phone company, they aren't culpable for bullying on their site, just as Verizon isn't if you bully someone over text.*" A few participants reported it was not social media companies' responsibility but clarified some caveats: "*It's not the social media's companies responsibility. You know the risk you're taking signing up for social media (as long as it's not hate speech / discriminatory).*"

Some participants, about 9%, believe social media companies do not care about achieving a fair resolution. Among those, most explained that they believed companies were motivated by profit rather than users. One said: "*None of them because they don't care about their users they just care how many users are on their platform that's why companies that have anonymous messages exist—anyone with any common sense knows they harm they will cause but they also know that people will use that platform*." Another participant similarly said: "*None. Social media companies are businesses which only work for increased revenue, not well-being of users*."

However, many participants (20%) reported that they trusted social media companies. One participant felt that social media companies have been recently improving their community guidelines for handling problems related to harassment. Another said: "*I would trust all the major companies such as Facebook, Twitter, Google, and the companies owned by or affiliated with them because [they] have teams for dealing with [this], clear in-app reporting, and easily access[ible] support.*"

Among individual social media companies, the most trusted site was Instagram, with 15% of participants stating that they trusted it to achieve a fair resolution for them, versus 1% saying they distrust it. Participants were optimistic about Instagram, saying "*Instagram seems well-regulated and it's easy to block/report someone*" and "*I think Instagram is one of the best social media sites for deleting wrong content. Ingstagram has few options for commenting, whereas Facebook has several areas (status', wall posts, pictures, groups, etc) that give a platform for bullying/harassment.*" Facebook was the second most trusted with 10% trust and 1% distrust and Twitter followed at 6% trust. One participant reported: "*Facebook actually, they tend to follow up on things that I report to tell me if they were taken down.*" Another similarly said: "*I would say Facebook because I feel they are very proactive in making sure content is appropriate. I also know firsthand they take fast action when something is reported.*" The other companies referenced—Snapchat, YouTube, TikTok, and Reddit—each had fewer than 3% of participants expressing distrust or trust. One of the Reddit supporters stated: "*Reddit is the most reliable platform I have seen. Not grounds for total chaos like 4chan, while not so primarily concerned about the bottom line like Facebook or Twitter. Their terms of agreement are pretty clear, and there is a structure that can keep corrupt sub-forum moderators from leaking their corruption elsewhere in the website.*"

Youth trust in institutions may vary by individual and group differences. We ran binary logistic regression models to explore what characteristics predict trust (see Table 1). Results suggest that older participants are more likely to trust Facebook. For example, one 24 year old participant said: "*I have noticed that Facebook will remove comments, and I assume other platforms would as well, but I don't know of any personally that have so I am not totally aure [sure]*." In contrast, for example, one 14 year-old participant said: "*Snapchat because when you report someone they actually ask what the problem is.*"

Participants who are transgender are more likely to distrust all companies. One participant who was transgender stated, "*People harassed me in high school over my gender and sexual identity"* and that *"Social media doesn't do shit for bullying.*" Another similarly said: "*when i came out as trans i was harassed on instagram pretty frequently, and i had people make posts that ruined my reputation on tumblr*" but that "*knowing they got in trouble for it*" helped them feel better. Perhaps surprisingly, participants on Free and Reduced Lunch programs are more likely to trust Facebook and Instagram and are less likely to not trust any companies; we return to this in the discussion.

4.2 Youth Preferences for Fair Resolutions

Among their reported preferences, youth were most likely to prefer an apology as a fair resolution to online harassment (62%). Women and participants with higher parent education levels were more likely to prefer apologies in general. One participant explained: "*The fairest thing would be any kind of apology. If they acknowledged what they did was wrong and they*

Table 1: Trust in Social Media Companies.

	Trust Facebook			Trust Instagram			Trust None		
	B(SE)	CI(L)	CI(U)	B(SE)	CI(L)	CI(U)	B(SE)	CI(L)	CI(U)
Intercept	-	-7.5	-4.2	-5.85(.84)***	-7.51	-4.20	-0.49(.24)*	-0.95	-0.02
age	0.16(.04)***	0.08	0.24	0.16(.04)***	0.08	0.24	-	-	-
woman	-	-	-	-	-	-	-	-	-
nonbinary	-	-	-	-	-	-	-	-	-
transgender	-	-	-	-	-	-	0.99(0.35)**	0.30	1.68
Asian	-	-	-	-	-	-	-	-	-
Black	-	-	-	-	-	-	-0.34(0.23)	-0.79	0.12
American Indian	-	-	-	-	-	-	-0.47(0.41)	-1.27	0.34
Hispanic	-0.65(.42)	-1.47	0.17	-0.65(0.42)	-1.48	0.17	-	-	-
Parent education	-	-	-	-	-	-	0.05(0.03)	-0.02	0.11
FRL	1.2(.25)***	0.72	1.69	1.2(.25)***	0.72	1.69	-0.36(0.17)*	-0.70	-0.02

[a] *p<.05, **p<.01, ***p<.001. B = beta; SE = standard error; CI(L) and CI(L) = confidence intervals at 2.5% and 97.5%. FRL = Free and Reduced Lunch program

regret it, that's the best possible outcome. If someone doesn't apologize, then depending on how bad their behavior was, banning them might be appropriate, but in an ideal world they would acknowledge what they did was wrong and apologize."

Twice as many participants preferred a private apology rather than a public apology (29% versus 14%). A chi-square test of independence showed that there was a significant difference between preferences for private versus public apologies, $X2$ (1, N = 830) = 51.59, $p < 0.001$. Among those who preferred the private apology, most preferred it to come via a direct message (27% of all participants) rather than an in-person apology (2% of all participants). Higher parent education level was associated with a preference for a public apology (see Table 2). For many participants, preferences for the private apology was associated with perceived authenticity of the apology, rather than merely virtue signaling via a public apology. A participant said: "*A private apology because to me that would show they're privately genuine and not putting on a show for the world to see*." Another noted that "*a personal apology via dm has helped a few times*" but that usually they would just tell the offender that they were wrong and then block them before they had a chance to reply. One participant wanted to see both an apology to the person via dm and a public apology explaining why they did what they did. The private apology was also preferred because as one said: "*I'm not about making my business public*." The apology channel was also aligned with what channel the harassment took place in: if it was public then the apology should be public, and vice versa. One participant suggested that a private apology and deleting posts was appropriate for less severe offenses; for more severe offenses, the offender should be banned and should give a public apology. They included statements like "*A public apology because they sought to humiliate me*" and "*Publicy apology since they destroyed your image*."

The next three preferred options were favored at a comparable level to each other, with between 24-30% of participants expressing a preference for deleting content, social support, and banning users. Deleting content refers to deleting the harassing post or having it removed, and was preferred by 30% of participants. These preferences included "*Delete post, act like it never happen in the first place and move on*" and "*I would love to have the hurtful posts removed and support from friends.*" Women and participants with higher parent education levels were more likely to express a preference for deleting content.

Social support involves support from friends, family, or other trusted people and was preferred by 25% of participants. A participant noted that: "*Support from friends would be much needed in a time like that. All I want to see from the other person is evidence that they've learned their mistake*." Another indicated that friends engaging as bystanders was important: "*Comments on the post from friends/strangers showing that it wasn't acceptable*." One expressed that other responses would not be effective whereas social support would be: "*I feel like when you ask for an apology, they're just writing what you want them to say and they don't really mean it. Deleted posts and banning isn't gonna fix it. I'd just ask friends for help*." Participants with higher parent education levels were more likely to express a preference for social support.

Banning harassers was supported by 24% of participants. One said: "Banning them. If you can't use social media correctly then you shouldn't be allowed to use it." Others were more comprehensive in their preferences, suggesting multiple different approaches as a response:

> *"Banning them for bullying, first things first. You don't want to fall victim again and at the same time you can prevent other people from being harassed. Apologies are always acceptable. The type depends on the location, like if you lived closer together you can meet and do a face to face apology, but if you live across the world a DM apology will work. Support from friends are always nice, since they will build you back up after horrific bullying tore you down to nothing but a shell of who you were."*

Some did not advocate for banning and instead wanted "*public shaming for the bully*." Another wanted to avoid public shaming, however, saying: "*Private [apology] I wouldn't like them getting bashed by people because they apologized*." One participant did not ask for banning but instead wanted behavior change from the offender, saying: "*Just stop being a lil bitch*." The binary logistic regression suggested that participants who were Asian or Black were less likely to support banning offenders. For example, one participant said: "*There needs to be deleted posts and also public support from friends. A public apology seems forced and banning them is too crazy like how everyone gets banned on twitter even for slight teasing*." Blocking was a preferred approach by 4% of participants which was sometimes preferred in place of other options and sometimes

Table 2: Preferences for fair resolutions.

	Private Apology			Ban Offender			Delete Content			Social Support		
	B(SE)	CI(L	CI(U)	B(SE)	CI(L)	CI(U)	B(SE)	CI(L)	CI(U)	B(SE)	CI(L	CI(U
Intercept	-	-	-0.46	-	-1.25	-0.61	-	-2.24	-1.31	-	-	-
Age	0.01(.03)	-	0.06	-	-	-	-	-	-	-	-	-
woman	-	-	-	-	-	-	0.58***(.16	0.26	0.90	-	-	-
nonbinary	-	-	-	-	-	-	-	-	-	-	-	-
transgender	-	-	-	-	-	-	0.66(.37)	-0.07	1.39	-0.99(.54)	-	0.06
Asian	-	-	-	-0.55(.25)*	-1.04	-0.07	-0.40(.22)	-0.83	0.04			
Black	-0.35(.25)	-	0.13	-0.64(.27)*	-1.17	-0.10	-	-	-	-	-	-
American Indian	-	-	-	-	-	-	-	-	-	-	-	-
Hispanic	-	-	-	-	-	-	-	-	-	-	-	-
parent education	0.08(.03)*	0.02	0.15	-0.25(.17)	-0.58	0.09	0.11(.03)**	0.04	0.17	0.07(.03)*	0.00	0.14
FRL	-	-	-	0.28(.18)	-0.06	0.63	-	-	-	-	-	-

[a] *p<.05, **p<.01, ***p<.001. B = beta; SE = standard error; CI(L) and CI(L) = confidence intervals at 2.5% and 97.5%; FRL = Free and Reduced Lunch program.

preferred in addition to them: "*Blocking them is also a quick and dirty solution. Banning seems unnecessary but maybe for repeat offenders*." Most indicated that it was low effort and expedient, since blocking did not require involvement of the offender or the platform.

5 DISCUSSION

This research explores how platforms can respond to online harassment in ways that support and acknowledge youth. Our results reveal that youth are more likely to distrust than trust social media platform responses to online harassment. They prefer apologies, primarily, and then deleting content, social support, and banning offenders, as responses to online harassment. Our results do not assess efficacy of these approaches; however, youth support for these approaches indicate opportunities for design exploration.

5.1 Punitive versus Restorative Approaches for Youth

Youth expressed enthusiasm for apologies after online harassment. They were more supportive of private apologies than public apologies, but both were supported in some contexts. This preference aligns with aspects of restorative justice approaches, which embrace accountability and repair [77]. Restorative justice asks that offenders be accountable for their actions and that targets feel harms have been repaired. Resolution and respect may not be attainable, however, and a restorative justice process does not need to lead to forgiveness or reconciliation [6, 10].

Though restorative justice is supposed to be healing for youth, in practice, it may not result in such outcomes. Not all restorative justice processes achieve their intended aims and the goals of youth restorative justice processes may be misaligned with their outcomes [14]. For example, an apology may be nongenuine or forced, which can magnify rather than reduce harm to youth [4, 30]. In our prior work with adults, transgender people did not like the idea of an apology, perhaps because it could come across as not genuine [65]. Prior research has also shown how transgender people experience targeted forms of abuse based on their identity, indicating that one-size-fits-all approaches may fail to support youth equitably [64, 65]. If an offender will not accept responsibility or will not repair harms, the justice process will not be successful. Thus, while apologies can be a conduit for justice, the delivery of an apology should not create an expectation of forgiveness from the target, nor should it imply that accountability was present.

An open question is how to decide whether to enact punitive (e.g. deleting content, banning users) versus reparative approaches (e.g. accountability, apologies) after online harassment. That is, which approach is better, and for whom, and who decides? Restorative justice approaches often involve mediation between offenders and victims. These have been effective in communities with preexisting ties, such as Indigenous communities [23, 28, 45, 46]. They have also been effective in youth contexts [10], though requiring youth to engage in mediation may be merely an alternative punishment rather than a lack of punishment. It is possible there are some online contexts where mediation could be fruitful: offenders and targets may not need to meet in-person, but can instead have a mediated online interaction where the target is buffered from further harm through the mediation process [14, 58]. An online mediated restorative justice process could be synchronous or asynchronous, video or text only, and could facilitate de-escalation using design friction techniques such as time delays for posting. Social media platforms could also incorporate trauma-informed approaches to online mediation that include trustworthiness, peer support, choice, and empowerment [78]. However, mediation processes may exacerbate harm, such as in gender-based violence where engagement with the abuser only fuels more abuse [11, 22]. Thus, restorative justice offers some pathways for repairing youth harassment experiences, but it can also contribute to significant harm to youth, and should not be implemented universally.

5.2 Should Youth Trust Social Media Companies?

Youth were more likely not to trust social media companies than to trust them. This aligns with youths' lower trust with institutions in general, including police, businesses, and education systems [82]. Many people's trust in, or view towards, an institution is shaped by how they system is presented to them. People are more likely to comply with outcomes when they believe a process was fair. Perceived legitimacy is also important for increasing perceptions of fairness. For example, if a court system is presented as having an "atmosphere of confusion" and as "unprofessional" then youth began to delegitimize the entire justice system as a whole based on this atmosphere inside the courtroom [24]. Similarly, schools that are perceived as harsh and punitive can make students feel excluded and ignored [37]. Many participants expressed that they did not trust companies like Facebook because they perceived them as caring more about making money than taking care of users. If social media companies' responses to online harassment appear inconsistent, or driven by biased human workers or algorithms, youth may experience decreased trust in the platform's ability to deliver a just process or outcome. Aligned with procedural justice approaches, social media companies could disclose the process they use to make content moderation decisions so that youth have a better understanding of the companies' values, which may increase their confidence in companies' ability to govern harassing behaviors.

Youth are more likely than older adults to have low confidence in other people's civic engagement, including other people's ability to respect the rights of people not like them, to treat others with respect, and to help others in needed [82]. Youth with low trust also tend to have lost confidence in other people, believing that everybody has become less reliable over time and this in turn makes it difficult to solve many of the country's problems [82]. This lack of interpersonal and institutional distrust may make it more difficult to mediate harassment between social media users. The inability to trust the process, or the people, makes it difficult to remediate harm and improve behavior [2, 25]. Recent scholarship has called for procedural justice approaches to moderating online content, arguing that transparency into how decisions are made can increase users' trust in and acceptance of those decisions [35, 70]. It may be possible to design responses to online harassment that are youth-oriented, such as designing conflict-resolution, de-escalation, and reparation into the moderation process [10, 46].

While procedural justice may support compliance with rules, we should consider whether we *want* youth to better comply with such rules. Trustworthiness is an antecedent of trust [13], and levels of trust in technology companies has declined over the past five years [16]. Trustworthiness relates to motivations for acting, and is often perceived to be a moral assessment [29]. Trustworthiness may also relate to credibility, authenticity, or reliability. Social media companies have faced public critique and regulation related to concerns about privacy, data management, and algorithmic bias [12, 19, 71]. Despite these critiques, youth still use social media platforms, suggesting a kind of trust paradox—akin to the oft-described privacy paradox [3]—in which individuals' expressed intentions are misaligned with actual behavior. However, these paradoxes overlook the costs of opting out, which can compromise individuals' social, professional, and economic opportunities [68]. They also overlook civic responsibilities to stay on platforms to engage with and critique the impact of those technologies on society, such as youth political activism on TikTok [75].

5.3 One Size Does Not Fit All

Youth may feel bound to institutions due to their reliance on them as they grow into the expectations of adulthood. Youth are exposed to institutions at early ages, such as education systems, health care systems, and juvenile justice systems [7, 24, 31]. While social media companies have sometimes responded to regulation with youth-focused designs, such as reducing collection of sensitive data and restricting the types of advertisements youth are exposed to, much of the social experience on social media platforms is the same as adults. Platforms could

implement youth-centered models that move beyond single-axis approaches like content or account removals, which do not offer opportunities for remediation or education. Platforms could also adopt harm-centered governance approaches using trauma-informed practices. Social workers and community workers use trauma-informed approaches to acknowledge existing trauma and to minimize the risk of retraumatizing during provision of services [40, 56]. A critical component of trauma-informed practice is that each person should be treated as an individual with unique experiences and needs, and that people will trust people and institutions who are trustworthy. Our results suggested that participants who were Black or Asian did not like the idea of banning offenders who had harassed others. Prior work found that Native American participants did not like the idea of banning, and suggested that may be because of that group's history of being removed forcibly from their communities [65]. It may be that our participants felt a similar opposition to what could result in unfair banning that is disproportionately experienced by minoritized communities. Though we did not measure preferred alternatives among our participants, prior work suggests that Black and Asian people like the idea of educating people about their identity as a response to online harassment [65].

Some of our results corroborated prior narratives, such as transgender people not trusting social media companies which have a history of harming those groups [27]. Our finding that older participants (young 20s) trust Facebook more may be explained by their greater presence and use of Facebook as compared to adolescents who tend to adopt new emerging technologies [41]. However, we also found some unanticipated results, including that participants on Free and Reduced Lunch programs (i.e. participants from lower income families) are more likely to trust Facebook and Instagram and are less likely to not trust any companies. Working class youth and their families tend to be harmed by institutions in ways that can lead to distrust (e.g. in healthcare [59]); it is possible our results would not be confirmed with additional research, or that there are other potential explanations for this population of youth to have higher trust in social media companies.

5.4 Limitations and Future Work

Our work centers youth wellbeing by inviting perspectives and experiences from youth themselves. Our work also falls in line with a growing movement towards justice as an orienting beacon for creating equitable and inclusive experiences online. Our text message method allowed us to capture responses from adolescents and young adults using a medium they are comfortable with; however, text message constrains the number of questions that can be asked and precludes following up in depth. For example, the first text-message question did not prompt participants about blocking behaviors, and blocking may have been more prominent in responses if it had been in the question prompt. We also did not ask about social media use and could not control for use in regression models; future work could expand this approach by pairing use with trust/distrust measures among youth. Our statistical analysis was exploratory in nature, and offers preliminary insights for subsequent hypothesis-testing as well as for measuring potential confounds like social media use, propensity to trust, and other measures. This work also did not examine how to implement preferred approaches. An important next step from a youth advocacy perspective would be to interview youth or engage them in design activities that explore how platforms might better support them.

This work reflects a U.S.-centered perspective and overlooks youth in many other countries around the world who are also exposed to widespread harassment on social media. Finally, and importantly, our sample was designed to reflect a broadly representative sample of US youth; however, youth from minoritized backgrounds may be more likely to be exposed to more harassment and more severe harassment; work could focus specifically on those populations to better support them.

6 CONCLUSION

Youth experience widespread online harassment. This work reveals that youth are more likely to distrust than trust social media companies to respond fairly to online harassment. It also reveals preferences for apologies from the offender, in addition to preferences for deleting harassing content, banning users who post harassing content, and social support. This work contributes to a nascent but expanding conversation around the limitations of extant criminal justice models centered on deleting content and banning users. It reflects on procedural justice processes that encourage institutional trust, and critically reflects on whether youth *should* trust social media companies. It draws attention to restorative justice principles, which have been implemented in youth contexts offline, to center accountability and repair after online harassment. Finally, it shed lights on individual differences in preferences and attitudes, highlighting that one size does not fit all when supporting youth who experience online harassment.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grants #1763297 and 1552503. We thank the MyVoice team and volunteers who reviewed the survey questions. We thank the participants who shared their experiences and perspectives with us.

REFERENCES

[1] Asad, M. 2019. Prefigurative Design as a Method for Research Justice. *Proceedings of the ACM on Human-Computer Interaction*. 3, CSCW (Nov. 2019), 200:1-200:18. DOI:https://doi.org/10.1145/3359302.

[2] Bailey, P.E., Slessor, G., Rieger, M., Rendell, P.G., Moustafa, A.A. and Ruffman, T. 2015. Trust and trustworthiness in young and older adults. *Psychology and Aging*. 30, 4 (2015), 977–986. DOI:https://doi.org/10.1037/a0039736.

[3] Barth, S. and de Jong, M.D.T. 2017. The privacy paradox – Investigating discrepancies between expressed privacy concerns and actual online behavior – A systematic literature review. *Telematics and Informatics*. 34, 7 (Nov. 2017), 1038–1058. DOI:https://doi.org/10.1016/j.tele.2017.04.013.

[4] Battistella, E.L. 2014. *Sorry about that: The Language of Public Apology*. Oxford University Press.

[5] Bradford, B., Grisel, F., Meares, T.L., Owens, E., Pineda, B.L., Shapiro, J., Tyler, T.R. and Peterman, D.E. 2019. *Report Of The Facebook Data Transparency Advisory Group*. Yale Justice Collaboratory.

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

[7] Brestan, E.V. and Eyberg, S.M. Effective psychosocial treatments of conduct-disordered children and adolescents: 29 years, 82 studies, and 5,272 kids. *Journal of Clinical Child Psychology*. 27, 180–189.

[8] Carstensen, L.L., Fung, H.H. and Charles, S.T. 2003. Socioemotional Selectivity Theory and the Regulation of Emotion in the Second Half of Life. *Motivation and Emotion*. 27, 2 (Jun. 2003), 103–123. DOI:https://doi.org/10.1023/A:1024569803230.

[9] Chandrasekharan, E., Samory, M., Srinivasan, A. and Gilbert, E. 2017. The Bag of Communities: Identifying Abusive Behavior Online with Preexisting Internet Data. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2017), 3175–3187.

[10] Choi, J.J., Bazemore, G. and Gilbert, M.J. 2012. Review of research on victims' experiences in restorative justice: Implications for youth justice. *Children and Youth Services Review*. 34, 1 (Jan. 2012), 35–42. DOI:https://doi.org/10.1016/j.childyouth.2011.08.011.

[11] Citron, D.K. 2014. *Hate Crimes in Cyberspace*. Harvard University Press.

[12] Cohen, N. 2019. Zuckerberg Wants Facebook to Build a Mind-Reading Machine. *Wired*.

[13] Colquitt, J.A., Scott, B.A. and LePine, J.A. 2007. Trust, trustworthiness, and trust propensity: A meta-analytic test of their unique relationships with risk taking and job performance. *Journal of Applied Psychology*. 92, 4 (2007), 909–927. DOI:https://doi.org/10.1037/0021-9010.92.4.909.

[14] Daly, K. and Tifft, L. 2007. The Limits of Restorative Justice. *Handbook of Restorative Justice: A Global Perspective*. Routledge.
[15] DeJonckheere, M., Nichols, L.P., Moniz, M.H., Sonneville, K.R., Vydiswaran, V.V., Zhao, X., Guetterman, T.C. and Chang, T. 2017. MyVoice National Text Message Survey of Youth Aged 14 to 24 Years: Study Protocol. *JMIR Research Protocols*. 6, 12 (2017), e247. DOI:https://doi.org/10.2196/resprot.8502.
[16] Doherty, C., Kiley, J. and Inquiries, D. 20036USA202-419-4300 | M.-857-8562 | F.-419-4372 | M. Americans have become much less positive about tech companies' impact on the U.S. *Pew Research Center*.
[17] Duggan, M. 2017. Online Harassment 2017. *Pew Research Center: Internet, Science & Tech*.
[18] Facebook forms a special ethics team to prevent bias in its A.I. software: 2018. *https://www.cnbc.com/2018/05/03/facebook-ethics-team-prevents-bias-in-ai-software.html*. Accessed: 2020-08-20.
[19] Facebook's algorithm bias only as neutral as their creators: *http://www.washingtontimes.com/news/2016/may/16/facebooks-algorithm-bias-only-as-neutral-as-their-/*. Accessed: 2016-09-19.
[20] Friedman, B., Khan, P.H. and Howe, D.C. 2000. Trust online. *Communications of the ACM*. 43, 12 (Dec. 2000), 34–40. DOI:https://doi.org/10.1145/355112.355120.
[21] Gillespie, T. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press.
[22] Goldberg, C. 2019. *Nobody's Victim: Fighting Psychos, Stalkers, Pervs, and Trolls*. Penguin.
[23] Graycar, A. and Grabosky, P. 2002. *The Cambridge Handbook of Australian Criminology*. Cambridge University Press.
[24] Greene, C., Sprott, J.B., Madon, N.S. and Jung, M. 2010. Punishing Processes in Youth Court: Procedural Justice, Court Atmosphere and Youths' Views of the Legitimacy of the Justice System1. *Canadian Journal of Criminology and Criminal Justice*. (Oct. 2010). DOI:https://doi.org/10.3138/cjccj.52.5.527.
[25] Growing to Trust: Evidence That Trust Increases and Sustains Well-Being Across the Life Span - Michael J. Poulin, Claudia M. Haase, 2015.
[26] Guiso, L., Sapienza, P. and Zingales, L. 2004. The Role of Social Capital in Financial Development. *American Economic Review*. 94, 3 (Jun. 2004), 526–556. DOI:https://doi.org/10.1257/0002828041464498.
[27] Haimson, O.L. and Hoffmann, A.L. 2016. Constructing and enforcing "authentic" identity online: Facebook, real names, and non-normative identities. *First Monday*. 21, 6 (Jun. 2016). DOI:https://doi.org/10.5210/fm.v21i6.6791.
[28] Hand, C.A., Hankes, J. and House, T. 2012. Restorative justice: the indigenous justice system. *Contemporary Justice Review*. 15, 4 (Dec. 2012), 449–467. DOI:https://doi.org/10.1080/10282580.2012.734576.
[29] Hardin, R. 2002. *Trust and Trustworthiness*. Russell Sage Foundation.
[30] Hayes, H. 2006. Apologies and Accounts in Youth Justice Conferencing: Reinterpreting Research Outcomes. *Contemporary Justice Review*. 9, 4 (Dec. 2006), 369–385. DOI:https://doi.org/10.1080/10282580601014292.
[31] Hinds, L. 2007. Building Police—Youth Relationships: The Importance of Procedural Justice. *Youth Justice*. 7, 3 (Dec. 2007), 195–209. DOI:https://doi.org/10.1177/1473225407082510.
[32] Hoffmann, A.L. 2019. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society*. 22, 7 (Jun. 2019), 900–915. DOI:https://doi.org/10.1080/1369118X.2019.1573912.
[33] Hosseini, H., Kannan, S., Zhang, B. and Poovendran, R. 2017. Deceiving Google's Perspective API Built for Detecting Toxic Comments. *arXiv:1702.08138 [cs]*. (Feb. 2017).
[34] Iyengar, V., Ghosh, D., Smith, T. and Krueger, and F. 2019. Age-Related Changes in Interpersonal Trust Behavior: Can Neuroscience Inform Public Policy? *NAM Perspectives*. (Jul. 2019). DOI:https://doi.org/10.31478/201906c.
[35] Jhaver, S., Bruckman, A. and Gilbert, E. 2019. Does Transparency in Moderation Really Matter?: User Behavior After Content Removal Explanations on Reddit. *Proc. ACM Hum.-Comput. Interact.* CSCW, 2 (Nov. 2019).

[36] Kaba, M. and Duda, J. 2018. Towards the horizon of abolition: A conversation with Mariame Kaba.
[37] Kupchik, A. 2016. *The Real School Safety Problem: The Long-Term Consequences of Harsh School Punishment*. Univ of California Press.
[38] Laloggia, J. and Inquiries 2019. *U.S. public has little confidence in social media companies to determine offensive content*. Pew Research Center.
[39] Lang, F.R. and Carstensen, L.L. 2002. Time counts: Future time perspective, goals, and social relationships. *Psychology and Aging*. 17, 1 (2002), 125–139. DOI:https://doi.org/10.1037/0882-7974.17.1.125.
[40] Levenson, J. 2017. Trauma-informed social work practice. *Social Work*. 62, 2 (2017), 105–113.
[41] Madden, M. Teens Haven't Abandoned Facebook (Yet). *Pew Research Center's Internet & American Life Project*.
[42] Mann, L., Harmoni, R. and Power, C. 1989. Adolescent decision-making: the development of competence. *Journal of Adolescence*. 12, 3 (Sep. 1989), 265–278. DOI:https://doi.org/10.1016/0140-1971(89)90077-8.
[43] McCaslin, W.D. 2013. *Justice As Healing: Indigenous Ways*. Living Justice Press.
[44] McDonald, N., Schoenebeck, S. and Forte, A. 2019. Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proceedings of ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '19)* (2019).
[45] Melton, A.P. 1995. Indigenous Justice Systems and Tribal Society Indian Tribal Courts and Justice: A Symposium. *Judicature*. 79, (1996 1995), 126–133.
[46] Meyer, J.F. 1998. History Repeats itself: Restorative Justice in Native American Communities. *Journal of Contemporary Criminal Justice*. 14, 1 (Feb. 1998), 42–57. DOI:https://doi.org/10.1177/1043986298014001004.
[47] Mingus, M. 2019. The Four Parts of Accountability: How To Give A Genuine Apology Part 1. *Leaving Evidence*.
[48] Murphy, K. 2015. Does procedural justice matter to youth? Comparing adults' and youths' willingness to collaborate with police. *Policing and Society*. 25, 1 (Jan. 2015), 53–76. DOI:https://doi.org/10.1080/10439463.2013.802786.
[49] Opinion | Could Restorative Justice Fix the Internet? - The New York Times: *https://www.nytimes.com/2019/08/20/opinion/internet-harassment-restorative-justice.html*. Accessed: 2021-01-06.
[50] Penner, E.K., Viljoen, J.L., Douglas, K.S. and Roesch, R. 2014. Procedural justice versus risk factors for offending: Predicting recidivism in youth. *Law and Human Behavior*. 38, 3 (2014), 225–237. DOI:https://doi.org/10.1037/lhb0000055.
[51] Pharo, H., Sim, C., Graham, M., Gross, J. and Hayne, H. 2011. Risky business: Executive function, personality, and reckless behavior during adolescence and emerging adulthood. *Behavioral Neuroscience*. 125, 6 (2011), 970–978. DOI:https://doi.org/10.1037/a0025768.
[52] Phillips, W. 2015. *This Is Why We Can't Have Nice Things: Mapping the Relationship Between Online Trolling and Mainstream Culture*. MIT Press.
[53] Prevent Cyberbullying: 2012. *https://www.stopbullying.gov/cyberbullying/prevention/index.html*. Accessed: 2019-08-23.
[54] Rainie, L., Anderson, J. and Albright, J. 2017. The Future of Free Speech, Trolls, Anonymity and Fake News Online. *Pew Research Center: Internet, Science & Tech*.
[55] Rainie, L., Keeter, S. and Perrin, A. 2019. Americans' Trust in Government, Each Other, Leaders. *Pew Research Center - U.S. Politics & Policy*.
[56] Raja, S., Hasnain, M., Hoersch, M., Gove-Yin, S. and Rajagopalan, C. 2015. Trauma Informed Care in Medicine. *Family & Community Health*. 38, 3 (Jul. 2015), 216–226. DOI:https://doi.org/10.1097/FCH.0000000000000071.
[57] ReCal for Ordinal, Interval, and Ratio Data (OIR) - dfreelon.org: *http://dfreelon.org/utils/recalfront/recal-oir*. Accessed: 2020-05-18.
[58] Regehr, C. and Gutheil, T. 2002. Apology, justice, and trauma recovery. *Journal of the American Academy of Psychiatry and the Law*. 30, 3 (2002), 425–430.
[59] Richardson, A., Allen, J.A., Xiao, H. and Vallone, D. 2012. Effects of Race/Ethnicity and Socioeconomic Status on Health Information-Seeking, Confidence, and Trust. *Journal of Health*

Care for the Poor and Underserved. 23, 4 (Oct. 2012), 1477–1493. DOI:https://doi.org/10.1353/hpu.2012.0181.

[60] Roberts, S.T. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press.

[61] Ronson, J. 2016. *So You've Been Publicly Shamed*. Penguin Publishing Group.

[62] Rossner, M. 2013. *Just emotions: rituals of restorative justice*. Oxford University Press.

[63] Hasinoff, A., Gibson, A.N., and Salehi, N. 2020. The promise of restorative justice in addressing online harm. *Brookings*.

[64] Scheuerman, M.K., Branham, S.M. and Hamidi, F. 2018. Safe Spaces and Safe Places: Unpacking Technology-Mediated Experiences of Safety and Harm with Transgender People. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (Nov. 2018), 155:1-155:27. DOI:https://doi.org/10.1145/3274424.

[65] Schoenebeck, S., Haimson, O. and Nakamura, L. 2020. Drawing from justice theories to support targets of online harassment. *New Media & Society*. (Mar. 2020), 1461444820913122. DOI:https://doi.org/10.1177/1461444820913122.

[66] Seering, J. 2020. Reconsidering Self-Moderation: the Role of Research in Supporting Community-Based Models for Online Content Moderation. *Proceedings of the ACM on Human-Computer Interaction*. 4, CSCW2 (Oct. 2020), 107:1-107:28. DOI:https://doi.org/10.1145/3415178.

[67] Smith, A. 2018. How Americans View Tech Companies. *Pew Research Center: Internet, Science & Tech*.

[68] Social Media Collective 2011. "If you don't like it, don't use it. It's that simple." ORLY? *Social Media Collective*.

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

[70] Suzor, N.P., West, S.M., Quodling, A. and York, J. 2019. What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation. *International Journal of Communication*. 13, 0 (Mar. 2019), 18.

[71] This is how much Americans trust Facebook, Google, Apple, and other big tech companies: 2020. *https://www.theverge.com/2020/3/2/21144680/verge-tech-survey-2020-trust-privacy-security-facebook-amazon-google-apple*. Accessed: 2020-05-18.

[72] Thomas, D.R. 2006. A General Inductive Approach for Analyzing Qualitative Evaluation Data. *American Journal of Evaluation*. 27, 2 (Jun. 2006), 237–246. DOI:https://doi.org/10.1177/1098214005283748.

[73] Tyler, T.R. 1988. What Is Procedural Justice - Criteria Used by Citizens to Assess the Fairness of Legal Procedures. *Law & Society Review*. 22, (1988), 103.

[74] United States Population Growth by Region: *https://www.census.gov/popclock/data_tables.php?component=growth*. Accessed: 2020-05-19.

[75] Upton-Clark, E. A Brief History of Pranks as Political Activism. *Teen Vogue*.

[76] U.S. Census Bureau QuickFacts: United States: *https://www.census.gov/quickfacts/fact/table/US/PST045219*. Accessed: 2020-05-19.

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

[78] What is Trauma-Informed Care? *http://socialwork.buffalo.edu/social-research/institutes-centers/institute-on-trauma-and-trauma-informed-care/what-is-trauma-informed-care.html*. Accessed: 2020-05-21.

[79] When It Comes to Abolition, Accountability Is a Gift: *https://www.bitchmedia.org/article/mariame-kaba-josie-duffy-rice-rethinking-accountability-abolition*. Accessed: 2021-01-06.

[80] Wohn, D.Y. 2019. Volunteer Moderators in Twitch Micro Communities: How They Get Involved, the Roles They Play, and the Emotional Labor They Experience. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2019), 160:1-160:13.

[81] Wulczyn, E., Thain, N. and Dixon, L. 2017. Ex Machina: Personal Attacks Seen at Scale. *Proceedings of the 26th International Conference on World Wide Web* (Republic and Canton of Geneva, Switzerland, 2017), 1391–1399.
[82] Younger Americans less trusting of other people, institutions | Pew Research Center: *https://www.pewresearch.org/fact-tank/2019/08/06/young-americans-are-less-trusting-of-other-people-and-key-institutions-than-their-elders/*. Accessed: 2020-05-18.

Received June 2020; revised October 2020; accepted December 2020.