When Online Harassment Is Perceived as Justified

Lindsay Blackwell, Tianying Chen, Sarita Schoenebeck, Cliff Lampe
University of Michigan School of Information
{lindsay.blackwell, cchent, sarita.schoenebeck, cacl}@umich.edu

Abstract

Most models of criminal justice seek to identify and punish offenders. However, these models break down in online environments, where offenders can hide behind anonymity and lagging legal systems. As a result, people turn to their own moral codes to sanction perceived offenses. Unfortunately, this vigilante justice is motivated by retribution, often resulting in personal attacks, public shaming, and doxing—behaviors known as online harassment. We conducted two online experiments (n=160; n=432) to test the relationship between retribution and the perception of online harassment as appropriate, justified, and deserved. Study 1 tested attitudes about online harassment when directed toward a woman who has stolen from an elderly couple. Study 2 tested the effects of social conformity and bystander intervention. We find that people believe online harassment is more deserved and more justified—but not more appropriate—when the target has committed some offense. Promisingly, we find that exposure to a bystander intervention reduces this perception. We discuss alternative approaches and designs for responding to harassment online.

Introduction

Online harassment refers to a broad spectrum of abusive behaviors enabled by technology platforms and used to target a specific user or users. This work is motivated by recent examples of harassment in online contexts that, although broadly viewed as harmful, are considered by some as justifiable responses to perceived social norm violations—a controversial form of social sanctioning. This "retributive harassment" can take many forms: high-profile examples include the 2013 public shaming of public relations executive Justine Sacco, the 2015 release of 40 million Ashley Madison users' personal and financial information, or the 2017 doxing of people who attended a white supremacist rally in Charlottesville, Virginia. Retributive harassment is especially widespread on social media sites such as Facebook and Twitter; however, why it happens and how to prevent it remain unknown.

Historically, abusive behavior online has been relegated to fringe cases—"narcissists, psychopaths, and sadists" (Buckels, Trapnell, and Paulhus 2014) who are either exceptions themselves, or inhabit atypical parts of the internet. Today, however, almost half of adult internet users in the U.S. have personally experienced online harassment, and a majority of users have witnessed others being harassed online (Duggan 2014; Duggan 2017; Lenhart et al. 2016; Rainie, Anderson, and Albright 2017). Although policies, reporting tools, and moderation strategies are improving (e.g., Perez 2017), most online platforms have failed to effectively curb harassing behaviors (Lenhart et al. 2016; Rainie, Anderson, and Albright 2017), and internet users and experts alike believe the problem is only getting worse (Rainie, Anderson, and Albright 2017).

This research aims to understand online harassment using a *retributive justice* framework. Retributive justice refers to a theory of punishment in which individuals who knowingly commit an act deemed to be morally wrong receive a proportional punishment for their misdeeds, sometimes referred to as "an eye for an eye" (Carlsmith and Darley 2008; Walen 2015). Retributive justice relies upon the assumption that everyday citizens possess intuitive judgments of "deservingness" that accurately and consistently express the degree of moral wrongdoing of others' acts. The integration of theories about justice and punishment with existing knowledge about social deviance and sanctioning has the potential to transform our current understanding of misbehavior in online spaces—in particular, when an instance of online harassment is perceived to be justified.

We conducted two online experiments to test the relationship between retributive justice and the perception of online harassment as justified or deserved. The first experiment tested whether exposure to a retributive prime—i.e., that the person being harassed had committed a crime—increases the belief that harassment is justified, deserved, or appropriate. The second experiment tested the effects of social influence on online harassment; specifically, whether conformity increases the belief that harassment is justified, deserved, or appropriate, and whether or not the presence of a bystander intervention would reduce these beliefs.

Investigating the relationship between orientations of justice and the perception of harassing behaviors online is an important step in better understanding what may motivate users to perpetrate online harassment—as well as what

motivates the decisions of moderators and bystanders, who may choose to take action (e.g., flagging or reporting) only against users whose actions they do *not* perceive to be justified. Ultimately, this research could generate a new understanding of social sanctioning online, influencing the design of technologies that support alternatives to retribution.

Related Work

Online harassment refers to a wide variety of abusive behaviors online, including but not limited to flaming (or the use of inflammatory language, name calling, or insults); doxing (or the public release of personally identifiable information, such as a home address or phone number); impersonation (or the use of another person's name or likeness without their consent); and public shaming (or visible humiliation intended to damage a person's reputation). These tactics are often employed concurrently, particularly when many individuals, acting collectively, target just one individual (sometimes referred to as "dogpiling").

Regulating Online Behavior

The first wave of Internet regulation, emerging in the 1980s (Rheingold 1993), involved establishing norms for good behavior and sometimes assigning community members special privileges (e.g., admins and moderators) to enforce those norms (Dibbell 1998; Kraut et al. 1996; Kraut et al. 2012; Lampe and Johnston 2005; Lampe et al. 2010). Regulation was also supported through moderation tools, such as reporting, flagging, and editorial rights (Matias et al. 2015; Pater et al. 2016). A second wave introduced crowdsourced approaches to regulation, such as the decentralized approaches used by Slashdot and Digg (Lampe and Resnick 2004; Poor 2005). These community moderation approaches have been effective in small online communities, such as LinuxChix, and sometimes in larger communities, such as Wikipedia (Bryant, Forte, and Bruckman 2005; Kraut et al. 2012; Panciera, Halfaker, and Terveen 2009); however, the size and scope of many online interactions have now outgrown normative regulation. The WELL had three thousand users in 1988 (Rheingold 1993); Twitter had 300 million monthly active users in July 2017 (Tsukayama, 2017).

Many emerging self-governance techniques in online communities, such as encouraging communities to establish their own rules (Matias 2017), cannot be implemented at scale. A more recent wave of regulation uses natural language processing and machine learning techniques to generate classifiers for detecting abusive language (Chandrasekharan et al. 2017; Hosseini et al. 2017; Wulczyn, Thain, and Dixon 2017; Yin et al. 2009). Though automated approaches have improved dramatically, they are subject to false positives and true negatives, with some harmful content eluding detection while other innocuous content is sanctioned (Hosseini et al. 2017). Furthermore, automatic detection efforts are relatively easy to bypass through subtle modification of language (Hosseini et al. 2017).

This work focuses on what we consider a fourth wave of regulation: everyday users enacting regulation by taking justice into their own hands. Many features of online interactions, such as anonymity, ephemerality, and persistence, are linked with impunity and freedom from "being held accountable for inappropriate online behaviour" (Diakopoulos and Naaman 2011; Hardaker 2010). Because offenders face little accountability for their actions online—and because legal systems are often unavailable or ineffective in online contexts—users have turned to forms of "vigilante justice" to enact punishments (Ronson 2015).

Certain affordances of online platforms, such as persistence, visibility, and broadcastability, may further enable this particular form of justice-seeking. On social media sites, users can easily capture and circulate content, even if the original author later deletes the post. Archived profile histories allow users to make character assessments quickly. When combined with the lack of affective cues in online contexts (Walther 1996), people's emotional arousal when faced with perceived injustice may lead them to rush to judgment and "fill in the blanks" about others they encounter online. Users can broadcast their desire for justice to wide audiences, and they can easily direct specific sanction requests to an offender's employer, family members, or other visible ties. Further, technological features such as likes, retweets and upvotes promote perceptions of endorsement—known as social proof—that can in turn lead to herd-like behaviors (Schultz et al. 2007; Steele, Spencer, and Aronson 2002). This has led to extreme and often disproportionate punishments for perceived offenses committed or circulated online, such as public shaming, physical threats, job termination, and sustained social isolation (Ronson 2015; Sydell 2017). Just as critically, these vigilante punishments can degrade civic discourse, promote disinformation, heighten polarization, and chill speech.

Justice and Retribution

In Kant's original conception of justice, the need for an institution to administer justice arises from the clear and immediate need to inflict proportionate suffering on an offending individual (Kant and Pluhar 1987). This philosophy is known as *retributive justice,* or the belief that offenders deserve sanctions that are proportional to the severity of their crimes. Retributivism is primarily preoccupied with delivering a 'just desert' for a morally wrong act (Kant and Pluhar 1987), sometimes referred to as "an eye for an eye" (Carlsmith and Darley 2008; Walen 2015). Retributive justice, unlike utilitarianism, highlights the need for proportionality in criminal sentencing (Wenzel et al. 2008). For example, in a retributive framework, the death penalty is considered a proportional punishment only for an offender who commits murder.

Retributivist intuitions of moral judgment interact with other theories of justice in complex ways. Carlsmith, Darley, and Robinson (2002) argue that even when individuals profess beliefs in the utilitarian-deterrence theory of justice (the belief that a punishment is just only if it effectively discourages others from committing the same crime), they nonetheless continue to apply retributivist assessments to punishment, judging offenders based on degree of moral wrongdoing (Carlsmith, Darley, and Robinson 2002). Moral judgment plays a powerful role in retribution and shapes cultural attitudes, policy, and law around appropriate punishments (Giner-Sorolla et al. 2012; Prinz 2007).

Retributive justice exists within particular social and institutional boundaries, and thus the parameters for what merits retributive punishment are socially constructed and contextual. Indeed, most formal justice systems consider intent when determining punishments. However, different cultures around the world—and even different states in the U.S.—have widely varied beliefs about the appropriateness of some punishments (e.g., death) for criminal offenses. On social media sites, users may seek retribution but have little guidance as to how to enact punishments, or even what an appropriate punishment may be. A widely-known example is that of Justine Sacco, who posted a racist tweet to her 170 followers while boarding a plane to South Africa (Ronson 2015). Her tweet was captured by mainstream media and resulted in threats of physical and sexual violence and (successful) demands that she be fired. By the time Sacco's flight had landed, the hashtag #HasJustineLandedYet was trending globally on Twitter.

This research seeks to better understand and intervene in online harassment by bridging theories of justice and the underlying circumstances that motivate users to participate in harassing behaviors online. To test the effect of an offense on people's perception of online harassment, we first hypothesize that:

H1: Exposure to a retributive prime increases belief that online harassment is a) justified; b) deserved; and c) appropriate.

Based on the principle of proportionality, or "an eye for an eye," we also hypothesize that participants will view online harassment as more justified and more deserved when the target's perceived offense is demonstrably greater. Second, we hypothesize that:

H2: Exposure to a larger retributive prime further increases belief that online harassment is a) justified; b) deserved; and c) appropriate.

In the Western world, punishment is enacted by a state or institution, and is typically designed to be fair and transparent in process. However, when people take justice into their own hands, it may reflect more individualistic traits and beliefs. Individual people have varied orientations toward retribution; thus, we hypothesize that:

H3: Propensity for retributive justice increases belief that online harassment is a) justified; b) deserved; and c) appropriate.

Social Norms and Conformity

Social norms—such as values, customs, stereotypes, and conventions—are "social frames of reference" that individuals first encounter through their interactions with others, and which later become internalized (Sherif 1936). Little is known about how and why norms emerge; however, the widely accepted instrumental theory posits that "norms tend to emerge to satisfy demands to mitigate negative externalities or to promote positive ones" (Hechter and Opp 2001). Thus, norms are most likely to emerge when they favorably impact a given community's goals (Opp 2001).

The perceived violation of a social norm is referred to as social deviance. Communities use deviance to establish boundaries—or rather, those who misbehave in turn establish community norms and how rules are made, enforced, and broken (Erikson 1964; Goode and Ben-Yehuda 2009). Communities develop norms for appropriateness and enforce those norms through sanctions, both formal (e.g., rules and laws) and informal (e.g., shame or ridicule).

Empirical evidence continues to suggest that group behavior influences individuals to behave similarly. Cialdini (2007) argues that descriptive norms offer "an information-processing advantage," in that by understanding how most people behave in a given situation, a social actor can more quickly decide how to behave themselves (Cialdini, Reno, and Kallgren 1990). Milgram, Bickman, and Berkowitz's 1969 experiment on the power of crowds is a classic example: when four people standing on a street corner look up at the sky, 80% of passersby will do the same. Normative appeals are most effective when individuals feel connected with a community or group—when we are uncertain about how to behave, we are more likely to "follow the herd," or conform to the perceived norms of a given social group (Goldstein, Cialdini, and Griskevicius 2008).

Conformity is a type of social influence in which changes in behavior or beliefs are motivated by a desire to adhere to the perceived social norms of a given group. A number of factors increase social conformity, including group size, group cohesiveness, status, self-esteem, and culture. This propensity toward social conformity facilitates distortions of perception (e.g., seeing objects or situations differently than they really are) and distortions of judgment (e.g., believing an act is okay only because other people appear to share that belief). In online environments, factors like relative anonymity, social distance, and social proof may also enhance disposition toward social conformity (Bogardus 1933; Cialdini 2001; Walther 1996). When people witness others engaging in a given behavior, they may seek to con-

form with the social norms of the group and engage in that behavior themselves. In the context of online harassment, the escalation of threats against a specific individual—sometimes referred to as 'dogpiling'—may be partially explained by the tendency to conform. We hypothesize that:

H4: Exposure to conformity increases belief that retributive online harassment is a) justified; b) deserved; and c) appropriate.

Bystander intervention

Bystander intervention is one potential antidote to undesirable social conformity. The concept of a bystander refers to a person who observes a situation and their subsequent decisions about whether or not to respond or intervene (Darley and Latané 1968). Intervening in an emergency situation can overcome what is called the bystander effect, where large groups of people observe but ignore offensive behaviors. There are several factors which contribute to the bystander effect, including ambiguity (particularly as emergency situations unfold) and diffusion of responsibility, or an individual's assumption that others are responsible for taking action (or have already done so). Empirical research confirms that the presence of bystanders in an emergency situation reduces helping responses (Fischer et al. 2011).

Promisingly, existing scholarship has also identified several factors that can reduce this bystander effect, including the perceived danger of an emergency, the bystander's relationship to the victim, and the potential risks associated with intervening (Fischer et al. 2011). While group behaviors promote conformity online, propensity toward conformity may be reduced when boundaries around appropriate behavior are questioned. We hypothesize that:

H5: Among conforming responses, exposure to bystander intervention decreases belief that retributive online harassment is a) justified; b) deserved; and c) appropriate.

Methods

We designed two experiments to test our hypotheses. Both studies were approved by an Institutional Review Board.

Recruitment

Participants for Study 1 were recruited through Twitter. For Study 2, participants were recruited via Twitter and Amazon Mechanical Turk (MTurk). During pilot testing, the survey took an average of 8 minutes to complete; thus, all study participants received $2 as compensation for their time, commensurate with a $15 hourly minimum wage.

Punishment Orientation Questionnaire

The Punishment Orientation Questionnaire (POQ) (Yamamoto 2014) is an 18-item scale developed to measure individual differences in punishment orientation. In both studies, a participant's score on the POQ's Harsh Retributive Scale (HRS) was used to operationally define their propensity for retributive justice.

Experiment 1: H1, H2, H3

The first study was a 3x1 between-subjects experiment with 3 parts and a total of 35 questions. The first part included a hypothetical scenario of harassment on Twitter and five questions to gauge participants' responses. We chose to simulate a tweet because of the ability for Twitter users to contact people outside of their immediate networks (unlike Facebook, for example, where most interactions occur between Facebook Friends), which would enable someone to engage in retributive harassment regardless of their relationship to the target. The second part of the survey contained the POQ (Yamamoto 2014), to assess participant's propensity for retributive justice. The final portion of the survey comprised twelve demographic questions (age, gender, race/ethnicity, etc.).

Participants were randomly assigned to one of three conditions: control, low-retributive prime, or high-retributive prime. Participants in the low-retributive prime condition were shown the following prime: "Sarah stole $100 from an elderly couple." Participants in the high-retributive prime were shown the same information, but with a higher theft amount: "Sarah stole $10,000 from an elderly couple." Participants in the control condition did not receive a prime. We chose not to include a prime in the control condition—instead of showing a "neutral" prime—because we did not believe a neutral interaction between Sarah (the harassment target) and an elderly couple was possible. In all conditions, participants were shown a harassing tweet sent by Amy to Sarah (see Figure 1). Names and avatars were meant to represent white women to control for any possible effects of race and gender.

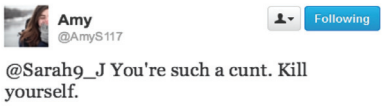

Figure 1. Simulated hostile tweet shown to participants.

Participants were asked to rate how *appropriate*, *deserved*, and *justified* Amy's tweet to Sarah was on a seven-item Likert scale from absolutely appropriate/deserved/justified to absolutely inappropriate/not deserved/not justified. Participants also responded to two open-ended questions: "If you saw this online, how would you feel?" and "If you saw this online, what (if anything) would you do?"

Experiment 2: H3, H4, H5

The second study also used a 3x1 between-subject design. Participants were randomly assigned into one of three conditions: control; conformity; and conformity + bystander intervention. Participants in all conditions were shown the following information: "Sarah stole $1,000 from an elderly couple." The survey used the same seven-item Likert scales

for appropriateness, deservedness, and justifiability used in study one, with three additional and original measures to understand how the participant would react in certain scenarios: a) "How likely would you be to call out Amy's behavior?" (seven-item Likert scale from extremely unlikely to extremely likely); b) "How likely would you be to call out Sarah's behavior?" (seven-item Likert scale from extremely unlikely to extremely likely); and c) "Whose behavior is more inappropriate?" (a seven-point sliding scale, with Sarah equal to 0 and Amy equal to 7).

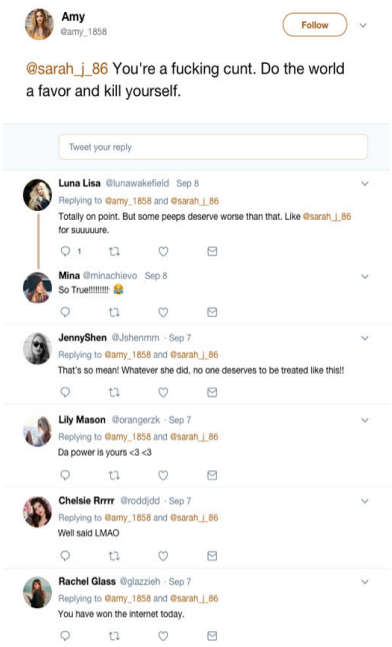

Figure 2. Simulated tweets shown to participants in the conformity + bystander intervention condition.

In each condition, participants were presented with a harassing tweet similar to the first study but with some adjusted content. We chose to change the "Following" text to Twitter's "Follow" button, to reduce ambiguity surfaced in the first study's open responses about whether participants knew Amy and Sarah. In the control condition, participants were only shown Amy's harassing tweet. In the conformity condition, participants were shown Amy's harassing tweet with conforming responses (i.e., responses supporting Amy's harassment of Sarah) from five other users. In the conformity + bystander intervention condition, participants saw Amy's harassing tweet with conforming responses from five other users, plus one user disagreeing with Amy's behavior (bystander intervention). We chose to add a sixth reply (see Figure 2) to avoid arbitrarily replacing one of the five replies used in the conformity condition. The conformity condition was otherwise identical to the conformity + bystander condition. As in study one, all display names and avatars were meant to represent white women, to control for any possible effects of race and gender.

Experiments 1 and 2: Open Responses

Both surveys included two open-ended questions: "If you saw this online, how would you feel?" and "If you saw this online, what (if anything) would you do?" We used an inductive approach to develop codes (Thomas 2006). The first author individually read through responses and noted codes by hand. After discussing these initial codes as a research team, we refined a list of codes (35 codes in total).

Resulting codes were organized around several themes, including but not limited to expressions of anger or disapproval toward Amy or Sarah; expressions of sympathy or understanding toward Amy or Sarah; feeling personally upset, offended, amused, or pleased; proportionality (overly harsh or insufficient punishment); expressing a desire to talk to Amy or Sarah, both privately and publicly; and specific actions participants would take if they saw this interaction in their feeds. Two researchers each coded several open responses to test and refine the codebook. In the first study, each open response was independently coded by two members of the research team. Because agreement was high, only the first author coded open responses from the second study. Quotations have been lightly edited for readability.

Participants

For study one, we received 541 total responses from Twitter. For study two, we received 597 total responses (150 responses from MTurk; 447 from Twitter). We removed invalid data from both studies using the following thresholds: a) incomplete responses (i.e., participants who did not reach the end of the survey); b) responses completed in under 200 seconds, which our pilot tests showed to be implausible; c) responses from duplicate IP addresses (all entries were removed); d) responses that had clearly identifiable spam (e.g., entering the word "good" for all open-response questions). For study one, a total of 160 valid cases remained after data cleaning (control group, n=56; low-retributive prime, n=49; high-retributive prime, n=55). For study two, a total of 432 valid cases (143 responses from MTurk; 289 from Twitter) remained after data cleaning (control group, n=145; conformity, n=146; conformity + bystander intervention, n=141).

Data analysis

We used SPSS and R for data cleaning and analysis, using a p-value of .05 for all statistical tests.

Study one: The dataset demonstrated a positively-skewed Poisson distribution, with a majority of the responses falling into either "absolutely inappropriate/not deserved/not justified" or "inappropriate/not deserved/not justified." Between-group one-way Welch's ANOVA was used to compare group mean between the three conditions to adjust for the violation of homogeneity of variance assumption of the

standard ANOVA test (Levene's test p <.0001). Similarly, we used a Games-Howell test for post-hoc multiple comparisons due to its robustness against violation of homogeneity of variance. Poisson regression was used to test the relationship between respondents' propensity for retributive justice and their responses (H3).

Study two: This dataset also demonstrated a positively-skewed Poisson distribution. Between-group one-way ANOVA and Tukey's HSD were used to compare means for deservedness. Between-group Welch's ANOVA and the Games-Howell post-hoc test were used for justifiability to adjust for the violation of the homogeneity of variance assumption of the standard ANOVA test (Levene's test p <.0001). Poisson regression was again used to test H3.

Results

Throughout, we use "offense" to describe the original offense committed by the harassment's target (Sarah's theft). We use "harassing tweet" to describe the retributive harassment targeting the offender (Amy's tweet).

Exposure and magnitude of retributive prime (H1, H2)

The first two hypotheses examined how participants' responses vary when presented with a retributive prime—in which the harassment's target (Sarah) has committed a prior offense (theft)—and whether this priming effect scales with the severity of the offense. H1 states that exposure to a retributive prime would increase the participant's belief that online harassment is justified.

Online harassment of an offender is justified and deserved, but not appropriate. In study one, a between-group one-way Welch's ANOVA revealed a significant difference across priming conditions for *deservedness* ($F(2, 92.887)=27.869$, $p<.001$) and *justifiability* ($F(2, 93.821)=15.115$, $p<.001$). No significant difference across priming conditions was found for *appropriateness* ($F(2, 103.942)=1.620$, $p=.203$). Further, Games-Howell post-hoc multiple comparison suggested that exposure to the retributive prime increased the participant's belief that harassment was *deserved* (M_{High}-$M_{Control}$ =1.864, SEM=.284, $p<.001$, $d=1.254$; M_{Low}-$M_{Control}$ =1.564, SEM=.313, $p<.001$, $d=1.013$) and *justified* (M_{High}-$M_{Control}$ =1.246, SEM=.261, $p<.001$, $d=.911$; M_{Low}-$M_{Control}$ =1.115, SEM=.292, $p<.001$, $d=.775$). In other words, H1 was partially supported: exposure to a retributive prime increases belief that online harassment is justified and deserved. That no significant difference was found for appropriateness suggests that even when Amy's harassment of Sarah was perceived as justified and deserved, participants still recognized that online harassment is not appropriate behavior.

H2 states that the belief of the justifiability of online harassment toward the offender should increase with the severity of the offense, consistent with the retributive value of proportionality. This hypothesis was not supported: no significant difference was found between the $100 and $10,000 primes. We specifically used theft as the offense because monetary amounts can be manipulated to be objectively higher or lower; however, it is possible that theft is perceived as a consistently offensive crime, regardless of the amount stolen. Future research should further test this hypothesis with different types of offenses, such as other types of crimes (e.g., vandalism or animal abuse) or social injustice (e.g., racism, white supremacy, or sexism).

Propensity for retributive justice (H3)

H3 states that propensity for retributive justice increases the belief that retributive harassment is justified, deserved, and appropriate. In both studies, a participant's score on the POQ's Harsh Retributive scale (HRS) was used to operationally define their propensity for retributive justice. We used Poisson regression to predict a participant's response to retributive harassment (Amy's tweet) based on their propensity for retributive justice and the priming condition (study one: control, low-retributive prime, and high-retributive prime; study two: control, conformity, and conformity + bystander intervention).

People who favor retributive justice find online harassment of an offender more deserved and more justified. Study one: A likelihood ratio test determined that the proposed model is significant for both *deservedness* ($\chi2=50.303$, $p<.001$) and *justifiability* ($\chi2=30.353$, $p<.001$), but not for *appropriateness*. For each one-point increase in the Harsh Retributive scale, there was a 2.9% increase in the participant's response to the *deservedness* of the harassing tweet (B=.029, Deviance=141.049, df=156, Wald $\chi2$ =4.751, exp(B)=1.029, p=.029) and a 3.6% increase in the participant's response to the *justifiability* of the harassing tweet (B=.036, Deviance=127.969, df=156, Wald $\chi2$ =6.324, exp(B)=1.037, p=.012). In other words, people who have a preference for retributive justice—commonly referred to an "eye for an eye"—believe that online harassment of an offender is more deserved and more justified (but not more appropriate) than do other people.

Study two: As expected (i.e., consistent with results from study one), the proposed model is significant for both *deservedness* ($\chi2=27.743$, $p<.001$) and *justifiability* ($\chi2=34.455$, $p<.001$). For each one-point increase in the Harsh Retributive scale, there was a 3.5% increase in the participant's response to the deservedness of the harassing tweet (B=.034, Deviance=529.535, df=428, Wald $\chi2$ =17.261, exp(B)=1.035, $p<.001$) and a 4.6% increase in the participant's response to the justifiability of the harassing tweet (B=.045, Deviance=482.377, df=428, Wald $\chi2$ =24.820, exp(B)=1.046, $p<.01$). We did observe small but significant differences in the mean scores between MTurk (n=143) and Twitter (n=289) responses for *appropriateness* (Twitter M=1.64, SD=1.08; MTurk M=2.42, SD=1.74), *deservedness* (Twitter M=3.15, SD=1.94; MTurk M=4.02, SD=2.20), and *justifiability* (Twitter M=2.54, SD=1.67;

MTurk M=3.53, SD=2.13). In other words, MTurk respondents perceived the harassing tweet as more appropriate, more deserved, and more justified than did Twitter respondents. This difference can be partially explained by MTurk respondents' higher scores on the Harsh Retributive Scale (Twitter M=14.24, SD=2.60; MTurk M=17.45, SD=3.07). Future research should assess a wider variety of participants to examine potentially meaningful differences in how users evaluate retributive harassment.

People who favor retributive justice are more likely to call out offensive behavior. In study two, participants were also asked how likely they would be to call out Amy's and Sarah's behavior. These were positively related ($r=.51$, $p<.001$), suggesting that people who reported being likely to call out Amy's behavior are also likely to call out Sarah's behavior. Further, participants' propensity for retributive justice was a significant predictor for both: for each one-point increase in the Harsh Retributive scale, there was a 2.3% increase in a participant's reported likelihood to call out Amy's behavior (retributive harassment). Similarly, for each one-point increase in the Harsh Retributive scale, there was a 2.5% increase in participant's reported likelihood to call out Sarah's behavior (theft). This indicates that people who favor retributive justice are more likely to voice public disapproval of offensive behavior.

Conformity and bystander intervention (H4, H5)

H4 and H5 examined the effects of social influence on a participant's perception of online harassment. H4 states that exposure to responses supporting the harassing tweet (i.e., conformity) increases the belief that retributive harassment is justified, deserved, and appropriate. No significant difference across priming conditions was found to support H4. This suggests that individuals' assessments of 'just deserts' may not be easily influenced by others.

Bystander interventions may help prevent dogpiling. H5—that exposure to bystander intervention among otherwise conforming responses should decrease belief that retributive harassment is justified, deserved, and appropriate—was partially supported. No significant effects were found for *appropriateness.* A between-group one-way ANOVA revealed a significant difference across priming conditions for *deservedness* ($F(2, 429)=4.247$, $p<.05$). Tukey's post-hoc comparison suggested that, compared to the control group, exposure to bystander intervention among conforming responses decreased the participant's belief that harassment was *deserved* ($M_{conformity+bystander} - M_{Control} = -.688$, $SEM=.243$, $p<.05$, $d=0.329$). A between-group one-way Welch's ANOVA also revealed a significant difference across priming conditions for *justifiability* ($F(2, 285.405)=4.220$, $p<.05$). Further Games-Howell post-hoc comparison suggested that compared to the control group, exposure to bystander intervention among other conforming responses decreased the participant's belief that harassment was *justified* ($M_{conformity+bystander} - M_{Control} = -.524$, $SEM=.214$, $p<.05$, $d=0.295$). In other words, bystander intervention reduces the perception of retributive harassment as justified or deserved.

In study two, participants were asked how likely they would be to call out Amy's and Sarah's behavior. In both cases, neither social conformity nor bystander intervention had a significant effect. Participants were also asked whose behavior was more inappropriate, using a seven-point sliding scale with Sarah equal to 0 and Amy equal to 7 (M=3, IQR=4). For each point increase in Amy's perceived inappropriateness, there was a 6.6% increase in participants' reported likelihood to call out Amy's behavior. However, we did not observe a corresponding effect on participants' likelihood to call out Sarah's behavior—suggesting that across all participants, retributive harassment merits public disapproval in a way theft does not.

Open responses

In both studies, participants were asked to respond to two open-ended questions: "If you saw this online, how would you feel?" and "If you saw this online, what (if anything) would you do?" Open responses are consistent with experimental results but add additional context for interpretation.

Context matters when determining just deserts. In study one's control condition (Amy's harassing tweet with no priming information about Sarah's offense), participants largely expressed that they would be personally upset or offended if they were to see this tweet online. Many participants identified Amy's behavior as being online harassment, which one respondent categorized as "not at all acceptable." However, even in the control condition, some participants expressed a desire to know more context, suggesting that there may be some situations in which Amy's behavior would be justified. Said one participant:

> "Telling someone to kill themselves is inappropriate in any circumstance. The rest [i.e., 'You're such a cunt'] depends on context, which is not available."

Some people said they would try to find out more about Sarah's offense from news websites, or would "read the comments to see how other people feel." Others said they would read through Sarah's or Amy's previous tweets to better assess their overall character.

Harassment can be a proportional punishment, but language matters. Across the low-retributive prime (Sarah stole $100 from an elderly couple) and high-retributive prime (Sarah stole $10,000 from an elderly couple) conditions in study one, participants assessed the proportionality of Amy's sanction based on Sarah's offense and Amy's choice of language. One respondent said that Sarah should have to face consequences for what she did, but that Amy went too far:

> "True, what Sarah did is terrible, and Sarah should have to face consequences for her act. But not death. The harsh judgement reminds me of the KKK and white supremacists, who believe their way is the only way."

Others agreed but maintained that Sarah should be called out for her behavior: "I feel that while confronting Sarah about stealing is the right thing to do, Amy shouldn't have insulted her in that way." Some participants, however, emphasized that "two wrongs don't make a right," and felt time would be better spent assisting the elderly victims of Sarah's crime. Other participants said they would contact the police or seek justice through other means. Said another respondent: "Sounds like Sarah's an asshole, but yelling garbage into the void does nothing to help the wronged."

Still, other participants—particularly in the high-retributive prime condition and in study two—were conflicted. One participant said "given that we know that Sarah stole," they "would not go to bat for her over a potentially over-the-line internet comment." Many participants said while they do not personally condone the language Amy used, they agree with what she said. Said one participant:

> "I think it's wrong to tell people to kill themselves (obviously), but who steals $10,000 from an elderly couple? She should be told in no uncertain terms what a horrible person she is."

Some participants were not at all conflicted by Amy's language, and said that if they were to see this tweet online, they would laugh, like or retweet Amy's tweet, or be otherwise amused. Others said they would "join in the bashing." One participant in study two applauded Amy for calling out Sarah's behavior: "At least some kids still have morals these days, even if they have foul mouths."

Online harassment has become normalized—and intervention is risky. Although some participants said they would report Amy's tweet for harassment or would message Sarah to offer emotional support, most participants said they would do nothing. One participant, when asked what (if anything) they would do, said: "Ignore it. It's not my battle." Other participants said they would react differently depending on whether or not they personally knew either of the women. Several participants who did not agree with Amy's behavior said they would not call her out or otherwise intervene, for fear of facing harassment themselves. Said one participant in study two: "In this day and age, I would be afraid to intervene." Another said: "Honestly, I probably wouldn't do anything—any direct response just opens you up to that kind of vitriol."

Across both studies, participants felt that online harassment was becoming normalized. One respondent said they don't feel it's ever appropriate to tell someone to kill themselves, but they felt desensitized to seeing these types of sentiments expressed online:

> "I honestly feel so desensitized to responses like Amy's. They are everywhere. I wouldn't feel much of anything—other than rolling my eyes and moving on."

Another respondent agreed: "It doesn't look like Amy is serious, so I'd shrug it off as typical Twitter hyperbole over Sarah's admittedly atrocious behavior." Participants directly associated this feeling of normalization with their unwillingness to report, call out, or otherwise intervene in Amy's harassment of Sarah. Said one participant: "I wouldn't do anything in particular. The internet is an awful place."

Discussion and Future Work

Designing technologies to encourage bystander action

Our results show that online harassment is perceived to be more justified and more deserved, but not more appropriate, when the target has committed some offense. Promisingly, exposure to a bystander intervention among other conforming responses decreased this perception—suggesting that designs encouraging bystander action could discourage harassment through normative enforcement.

Platforms can encourage bystanders to intervene by reducing ambiguity and diffusion of responsibility, factors which contribute to bystander apathy (Darley and Latané 1968). Indeed, recent research suggests that bystanders are motivated to intervene when they understand the breadth and impact of harassment, factors which are obscured in distributed, cue-sparse environments (Blackwell et al. 2017). Experimental research confirms that bystanders feel more personally responsible, and are more likely to intervene directly, when exposed to multiple instances of harassment targeting a single user (Kazerooni et al. 2018). Given these findings, social media platforms should counteract ambiguity by making the harmful impacts of online harassment more visible. Further, although many current interventions aim to obscure or hide harassment from both targets and bystanders (e.g., blocklists; block and mute tools; Twitter's "Quality Filter"), reminding potential bystanders that online abuse is both prevalent and inappropriate could foster a greater sense of personal responsibility.

Online bystanders are more likely to intervene in indirect ways (e.g., by reporting content to platforms) than by responding directly to perpetrators, due to the social and physical risks of direct intervention (Dillon and Bushman 2015). Platforms should prioritize simple, indirect interventions that do not put bystanders at risk. For example, HeartMob (iheartmob.org) provides bystanders with specific and private ways to take action, such as sending a supportive message or documenting abuse (Blackwell et al. 2017). Finally, some participants said they would look to other users' responses to determine how they themselves should act (i.e., descriptive norms); anonymously highlighting bystander interventions when they do occur may encourage other users to do the same.

Designing technologies to mitigate retribution

Our qualitative data suggests that some people censor themselves due to fear of retribution, suggesting that retributive harassment may contribute to chilled speech online. This could be particularly damaging for marginalized populations, including women, people of color, and LGBT people, who are already more likely to censor themselves online because they fear facing harassment (Duggan 2014; Lenhart et al. 2016; Rainie, Anderson, and Albright 2017). A danger of retributive harassment, and its widespread use, is that marginalized voices will be silenced while socially dominant perspectives are amplified.

Because the affordances of existing social media platforms exacerbate retributive harassment—and also limit potential consequences for those who choose to engage in vigilantism—we should instead consider designing platforms that encourage alternative forms of justice-seeking. An emerging alternative to retributive justice is *restorative justice,* which prioritizes improving society for the future. Restorative justice provides a voice to both victim and offender: the victim is encouraged to express a willingness to forgive, and the offender is encouraged to accept responsibility for their actions, with the goal of mending conflicts between individuals and communities (Wenzel et al. 2008).

Future work should explore ways of integrating restorative approaches into the design of online communities. Social media platforms could algorithmically detect surges of retributive harassment and experiment with designs that introduce mediation, reconciliation, and proportionality. This might involve the use of deescalating language that draws on shared experiences and understanding, mechanisms that enable or require social resolution, or the creation of spaces where communities can voice their feelings and concerns (Simonson and Staw 1992). For example, a new type of temporary Facebook Group could serve as a moderated platform for communities to work together with offenders to re-establish and validate relevant community values, restoring justice through social consensus (Wenzel et al. 2008). If social media platforms were to leverage their existing community features to encourage restorative mediation, justice could be restored without the use of retributive sanctions—promoting civil and inclusive participation online by enabling reconciliation at scale.

Conclusion

We propose the concept of *retributive harassment* to explain why and how certain kinds of online harassment occur—namely, when online harassment is used as a controversial form of social sanctioning. We reflect on the affordances of social media platforms that enable retributive harassment, and we advocate for the design of systems that encourage more restorative forms of justice-seeking.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant #1318143. We thank Zoë Wilkinson Saldaña for contributions to the research design.

References

Becker, H. S. 1963. *Outsiders: Studies in the Sociology of Deviance.* London: Free Press of Glencoe.

Blackwell, L.; Dimond, J.; Schoenebeck, S.; and Lampe, C. 2018. Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. In *Proceedings of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing* (CSCW '18). ACM.

Bogardus, E. S. 1933. A Social Distance Scale. Sociology & Social Research 17: 265–271.

Bryant, S. L.; Forte, A.; and Bruckman, A. 2005. Becoming Wikipedian: Transformation of Participation in a Collaborative Online Encyclopedia. In *Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work* (GROUP '05). ACM.

Buckels, E. E.; Trapnell, P. D.; and Paulhus, D. L. 2014. Trolls Just Want to Have Fun. *Personality and Individual Differences 67*, 97–102.

Carlsmith, K. M.; Darley, J. M.; and Robinson, P.H. 2002. Why Do We Punish? Deterrence and Just Deserts as Motives for Punishment. *Journal of Personality and Social Psychology 83*(2): 284–299.

Carlsmith, K. M., and Darley, J. M. 2008. Psychological Aspects of Retributive Justice. *Advances in Experimental Social Psychology 40:* 193–236.

Chandrasekharan, E.; Samory, M.; Srinivasan, A.; and Gilbert, E. 2017. The Bag of Communities: Identifying Abusive Behavior Online with Preexisting Internet Data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (CHI '17). ACM.

Cialdini, R. 2001. *Influence: Science and Practice.* Needham, MA: Allyn and Bacon.

Cialdini, R. B. 2007. Descriptive Social Norms as Underappreciated Sources of Social Control. *Psychometrika 72*(2): 263.

Cialdini, R. B.; Reno, R. R.; and Kallgren, C. A. 1990. A Focus Theory of Normative Conduct: Recycling the Concept of Norms to Reduce Littering in Public Places. *Journal of Personality and Social Psychology 58*(6): 1015.

Darley, J. M., and B. Latané. 1968. Bystander Intervention in Emergencies: Diffusion of Responsibility. *Journal of Personality and Social Psychology 8*(4): 377–383.

Diakopoulos, N., and Naaman, M. 2011. Towards Quality Discourse in Online News Comments. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work* (CSCW '11). ACM.

Dibbell, J. 1998. *My Tiny Life: Crime and Passion in a Virtual World.* Henry Holt & Company.

Dillon, K. P., and Bushman, B. J. 2015. Unresponsive or unnoticed?: Cyberbystander intervention in an experimental cyberbullying context. *Computers in Human Behavior 45:* 144–150.

Duggan, M.; Rainie, L.; Smith, A.; Funk, C.; Lenhart, A.; and Madden, M. 2014. Online Harassment. Pew Research Center.

Duggan, M., and Smith, A. 2017. Online Harassment 2017. Pew Research Center.

Fischer, P.; Krueger, J. I.; Greitemeyer, T.; Vogrincic, C.; Kastenmüller, A.; Frey, D.; Heene, M.; Wicher, M.; and Kainbacher, M. 2011. The Bystander-Effect: a Meta-Analytic Review on Bystander Intervention in Dangerous and Non-Dangerous Emergencies. *Psychological bulletin 137*(4): 517.

Giner-Sorolla, R.; Bosson, J.; Caswell, A.; and Hettinger, V. 2012. Emotions in Sexual Morality: Testing the Separate Elicitors of Anger and Disgust. *Cognition & Emotion 26*(7): 1208–1222.

Goldstein, N. J.; Cialdini, R. B.; and Griskevicius, V. 2008. A Room with a Viewpoint: Using Social Norms to Motivate Environmental Conservation in Hotels. *Journal of Consumer Research 35*(3): 472–482.

Goode, E., and Ben-Yehuda, N. 2009. *Moral Panics: The Social Construction of Deviance.* Wiley-Blackwell.

Hardaker, C. 2010. Trolling in Asynchronous Computer-Mediated Communication: From User Discussions to Academic Definitions. *Journal of Politeness Research 6*(2).

Hechter, M., and Opp, K. D. 2001. *Social Norms.* Russell Sage Foundation.

Hosseini, H.; Kannan, S.; Zhang, B.; and Poovendran, R. 2017. Deceiving Google's Perspective API Built for Detecting Toxic Comments. *arXiv preprint arXiv:1702.08138.*

Kant, I., and Pluhar, W. 1987. *Critique of Judgment.* Hackett Publishing.

Kazerooni, F.; Taylor, S. H.; Bazarova, N. N.; and Whitlock, J. 2018. Cyberbullying Bystander Intervention: The Number of Offenders and Retweeting Predict Likelihood of Helping a Cyberbullying Victim. *Journal of Computer-Mediated Communication.*

Kraut, R.; Scherlis, W.; Mukhopadhyay, T.; Manning, J.; and Kiesler, S. 1996. HomeNet: A Field Trial of Residential Internet Services. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 284-291. ACM.

Kraut, R. E.; Resnick, P.; Kiesler, S.; Burke, M.; Chen, Y.; Kittur, N.; Konstan, J.; Ren, Y.; and Riedl, J. 2012. *Building Successful Online Communities: Evidence-Based Social Design.* MIT Press.

Lampe, C., and Resnick, P. 2014. Slash(dot) and Burn: Distributed Moderation in a Large Online Conversation Space. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* pp. 543-550. ACM.

Lampe, C.; Wash, R.; Velasquez, A.; and Ozkaya, E. 2010. Motivations to Participate in Online Communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '10). ACM.

Lenhart, A.; Ybarra, M.; Zickuhr, K.; and Prive-Feeney, M. 2016. Online Harassment, Digital Abuse, and Cyberstalking in America. Data & Society Research Institute.

Matias, J. N. 2017. Posting Rules in Online Discussions Prevents Problems & Increases Participation. CivilServant.

Matias, J. N.; Johnson, A.; Boesel, W. E.; Keegan, B.; Friedman, J.; and DeTar, C. 2015. Reporting, Reviewing, and Responding to Harassment on Twitter. Women, Action, and the Media.

Milgram, S.; Bickman, L.; and Berkowitz, L. 1969. Note on the Drawing Power of Crowds of Different Size. *Journal of Personality and Social Psychology 13*(2): 79–82.

Opp, K. D. 2001. How Do Norms Emerge? An Outline of a Theory. *Mind & Society 2*(1): 101–128.

Panciera, K.; Halfaker, A.; and Terveen, L. 2009. Wikipedians Are Born, Not Made: A Study of Power Editors on Wikipedia. In *Proceedings of the ACM 2009 International Conference on Supporting Group Work* (GROUP '09). ACM.

Pater, J. A.; Kim, M. K.; Mynatt, E. D.; and Fiesler, C. 2016. Characterizations of Online Harassment: Comparing Policies Across Social Media Platforms. In *Proceedings of the 19th International Conference on Supporting Group Work* (GROUP '16). ACM.

Perez, S. 2017. Twitter Adds More Anti-Abuse Measures Focused on Banning Accounts, Silencing Bullying. *TechCrunch.*

Poor, N. 2005. Mechanisms of an Online Public Sphere: The Website Slashdot. *Journal of Computer-Mediated Communication 10*(2): 00–00.

Prinz, J. (2008). Is Morality Innate. *Moral psychology, 1,* 367-406.

Rainie, L.; Anderson, J.; and Albright, J. 2017. The Future of Free Speech, Trolls, Anonymity and Fake News Online. Pew Research Center.

Rheingold, H. 1993. *The Virtual Community: Homesteading on the Electronic Frontier.* New York: Harper Collins.

Ronson, J. 2015. How One Stupid Tweet Blew Up Justine Sacco's Life. *The New York Times.*

Schultz, P. W.; Nolan, J. M.; Cialdini, R. B.; Goldstein, N. J.; and Griskevicius, V. 2007. The Constructive, Destructive, and Reconstructive Power of Social Norms. *Psychological Science 18*(5): 429–434.

Sherif, M. 1936. *The Psychology of Social Norms.* Oxford, England: Harper.

Simonson, I., and Staw, B. M. 1992. Deescalation Strategies: A Comparison of Techniques for Reducing Commitment to Losing Courses of Action. *Journal of Applied Psychology 77*(4): 419–426.

Steele, C. M.; Spencer, S. J.; and Aronson, J. 2002. Contending with Group Image: The Psychology of Stereotype and Social Identity Threat. *Advances in Experimental and Social Psychology* 14: 379–407.

Sydell, L. 2017. Kyle Quinn Hid At A Friend's House After Being Misidentified On Twitter As A Racist. *NPR.*

Thomas, D. R. 2006. A General Inductive Approach for Analyzing Qualitative Evaluation Data. *American Journal of Evaluation 27*(2): 237–246.

Tsukayama, H. 2017. Twitter Lost 2 Million Users in the U.S. Last Quarter. *The Washington Post.*

Walen, A. D. 2015. Proof Beyond a Reasonable Doubt: A Balances Retributive Account. *La. L. Rev., 76,* 355.

Walther, J. B. 1996. Computer-Mediated Communication: Impersonal, Interpersonal, and Hyperpersonal Interaction. *Communication Research 23*(1): 3–43.

Wenzel, M.; Okimoto, T. G.; Feather, N. T.; and Platow, M. J. 2008. Retributive and Restorative Justice. *Law and Human Behavior 32*(5): 375–389.

Wulczyn, E.; Thain, N.; and Dixon, L. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th International Conference on World Wide Web* (WWW '17).

Yamamoto, S. 2014. The Reasons We Punish: Creating and Validating a Measure of Utilitarian and Retributive Punishment Orientation. Carleton University.

Yin, D.; Xue, Z.; Hong, L.; Davison, B. D.; Kontostathis, A.; and Edwards, L. 2009. Detection of Harassment on Web 2.0. *Proceedings of the Content Analysis in the WEB,* 2, 1-7.