

# Morally Motivated Networked Harassment as Normative Reinforcement

Alice E. Marwick 

Social Media + Society  
April-June 2021: 1–13  
© The Author(s) 2021  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/20563051211021378  
journals.sagepub.com/home/sms  


## Abstract

While online harassment is recognized as a significant problem, most scholarship focuses on descriptions of harassment and its effects. We lack explanations of *why* people engage in online harassment beyond simple bias or dislike. This article puts forth an explanatory model where networked harassment on social media functions as a mechanism to enforce social order. Drawing from examples of networked harassment taken from qualitative interviews with people who have experienced harassment ( $n=28$ ) and Trust & Safety workers at social platforms ( $n=9$ ), the article builds on Brady, Crockett, and Bavel's model of moral contagion to explore how moral outrage is used to justify networked harassment on social media. In morally motivated networked harassment, a member of a social network or online community accuses a target of violating their network's norms, triggering moral outrage. Network members send harassing messages to the target, reinforcing their adherence to the norm and signaling network membership. Frequently, harassment results in the accused self-censoring and thus regulates speech on social media. Neither platforms nor legal regulations protect against this form of harassment. This model explains why people participate in networked harassment and suggests possible interventions to decrease its prevalence.

## Keywords

networked harassment, social norms, amplification, morality, networked audience

## Introduction

Elise is an Asian American musician in her twenties, active on Twitter. In the spring of 2019, many people in Elise's Twitter network posted tweets criticizing a White-owned Chinese restaurant that promoted itself as a "clean" alternative to "unhealthy" Chinese food. Elise added to the criticism, posting a Twitter thread that discussed racist stereotypes and the history of Chinese American food. Her tweets went viral. She received hundreds of attacks over a period of days, including angry tweets, comments on her blogs and YouTube videos, and threatening emails.<sup>1</sup> Most of the harassing tweets framed Elise as a racist. For example, one tweet read,

Being a racist for the sake of being a racist is disgusting. These people are literally just trying to earn a living and you are single handedly trying to put them out of business, and for what. For what does this gain you? People eat better food, you get nothing in return.

This tweet does not portray Elise as an Asian American activist protesting a restaurant for its racism, but as an anti-White racist trying to put a hardworking female restaurant owner out of business. The continuity of the attacks on Elise

show that her accusers were able to reframe her behavior as immoral and even threatening, thus justifying harassing her. In this case, the network harassing her reinforced their belief that people of color calling out White racism is equivalent to or worse than White racism itself.

While harassment takes many forms, this article seeks to understand *networked harassment*, in which an individual is harassed by a group of people networked through social media. While most research describes the prevalence of harassing incidents and their impact on those who experience them, we lack an understanding of why harassment takes place. This article puts forth an explanatory model in which networked harassment functions as a mechanism to enforce social order. Drawing from a set of qualitative

The University of North Carolina at Chapel Hill, USA

### Corresponding Author:

Alice E. Marwick, Department of Communication, Center for Information, Technology, and Public Life, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3285, USA.  
Email: amarwick@unc.edu  
Twitter: @alictiara



interviews with people who have experienced harassment ( $n=28$ ) and workers at Trust & Safety platforms ( $n=9$ ), the article traces how moral outrage is used to justify networked harassment, which then functions to reinforce social norms and sanction violating behavior. In the example above, Elise's first tweet reflects her own moral beliefs that the restaurant furthered racist stereotypes about Chinese-Americans, while the harassment she received reframed her criticism as anti-White "reverse racism." This demonstrates that harassment is a tactic used by people across the ideological spectrum. However, due to fundamental power differentials that privilege Whiteness and maleness, those who challenge these structures are more likely to face harassment, systematically removing minority voices from the public sphere.

In this model, which I call *morally motivated networked harassment* (MMNH), a member of a social network or online community accuses an individual (less commonly a brand or organization) of violating the networks' moral norms. Frequently, the accusation is amplified by a highly followed network node, triggering moral outrage throughout the networked audience. Members of the network send harassing messages to the individual, reinforcing their own adherence to the norm and signaling network membership, thus re-inscribing the norm and reinforcing the network's values. The accusation is often escalated by the networked audience, fueling moral outrage and justifying further harassment. Frequently, harassment results in the accused self-censoring. As a result, networked harassment becomes a regulating force for speech on social media. This article outlines this model and its consequences.

## Online Harassment

"Online harassment" is an umbrella term widely used across fields to encompass a variety of behaviors. The term was first used by cyberbullying scholars to mean "bullying" (Tokunaga, 2010), "rude or mean comments, or spreading of rumors" (Ybarra & Mitchell, 2008), or "threats or other offensive behavior . . . sent online to the youth or posted online about the youth for others to see" (Finkelhor et al., 2000). This scholarship primarily concerned children and teenagers, where harasser and harassed knew each other, were the same age and lived in the same general location, and harassment took place in front of one's peer group off or online.

More recently, scholars have conceptualized "harassment" more expansively. Lenhart et al. identified 10 types of harassment, including physical threats, name-calling, impersonation, spreading rumors, and encouraging others to harass a target (Lenhart et al., 2016). A 2017 Pew report similarly included offensive name-calling, intentional embarrassment, physical threats, stalking, persistent harassment over time, and sexual harassment in its definition of harassment (Duggan, 2017). The broadness of these definitions suggests that anything from a single instance of name-calling to persistent, serious abuse can be labeled harassment. In response,

my previous research furthered a taxonomy of online harassment distinguishing between instances of *dyadic harassment*, when one person harasses another, resembling the dynamics of stalking or sexual violence; *normalized harassment*, in which name-calling or insults are common in online spaces like networked gaming; *networked harassment*, in which an individual is harassed by a group of people connected by social media; and more nebulous interpersonal situations in which at least one participant strategically labels an incident or set of incidents as "harassment" although it may not meet evaluative criteria as such (Marwick, in press).

Importantly, most work on harassment has been done by feminist scholars, since online harassment is more common for women and nonbinary people, particularly women of color, queer women, and women in the public eye (Barton & Storm, 2014; Krook, 2017; Lenhart et al., 2016; Sobieraj, 2020; Vitak et al., 2017). Such scholarship shows that harassment is often used to police women's online behavior and may have a chilling effect on women's participation in the public sphere both on and offline. As a result, scholars deploy terms like "online hate," "e-bile," "gender trolling," and "online misogyny" to connect online behavior such as "revenge porn" and digitally enabled sexual violence to structural sexism and violence against women (Banet-Weiser & Miltner, 2016; Citron, 2014; Eikren & Ingram-Waters, 2016; Henry & Powell, 2015; Jane, 2014; Mantilla, 2013; McGlynn et al., 2017). However, while most of this research argues that online harassment is caused by misogyny, there exist no explanatory models of *why* misogyny results in networked harassment.

I build on this feminist scholarship to connect networked harassment and social shaming. I am particularly interested in understanding how networked harassment functions as a method of social shaming which, as Kate Klonick (2015) writes, "involves the attempt to enforce either a real, or perceived, violation of a social norm." Shaming is thus a form of public moral criticism which serves to uphold social norms through stigma and humiliation (Billingham & Parr, 2020; Nussbaum, 2009). In contrast to the previous literature, my research shows that networked harassment takes place across ideological boundaries and is used by members of left-leaning and nonpolitical networks as well as those on the right. However, people who challenge normative power structures (such as feminists, anti-racist activists, gender non-conforming, and LGBTQ+ [lesbian, gay, bisexual, transgender, and queer] people) are more likely to be harassed by people who adhere to traditional social norms which privilege Whiteness, heteronormativity, maleness, and so forth.

## Moral Outrage, Boundaries, and Justifications for Harassment

Previous research demonstrates that perpetrators of harassment often believe their actions are justified (Blackwell et al., 2018; Jhaver et al., 2018). For example, a Minnesota

dentist named Walter Palmer killed a lion in a trophy hunt in Zimbabwe. When this was publicized on social media, he received worldwide harassment, including death threats and “lion killer” spray-painted on his house. This anger represented real outrage about big game hunting, wildlife conservation, and American arrogance, meaning that people who participated in the harassment believed they were in the right. However legally Palmer had done nothing wrong, and he is certainly not the only person to participate in such big-game hunting (Anderson, 2018). In many cases, an accusation of a norm violation spreads through a network whose members share a moral basis for the justification (Lewis et al., 2021). For instance, a progressive activist advocating deplatforming alt-right influencers might be labeled “anti-free speech” or “censoring” by right-wing network participants, while members of a left-wing network view them differently. In this case, the two networks have different priorities, values, and community norms. However, the context collapse endemic to large social platforms allows for networks with radically different norms and mores to be visible to each other (Marwick & boyd, 2011).

The renowned sociologist Michele Lamont defined *symbolic boundaries* as “distinctions that groups create between one another” (Lamont, 2017, p. 14). Her research shows that symbolic boundaries are often drawn based on real or imagined moral criteria. She interviewed middle-class Americans, noting that they harshly judged poor people as lazy and tasteless. In contrast, members of the French working-class saw the upper-middle class as selfish and narcissistic. Lamont’s model of boundary formation argues that such moral and ethical boundaries between groups are constituted through meaning-making, and that individual-level enforcement of symbolic boundaries helps to create and maintain cultural, institutional, and social differences (Lamont & Molnár, 2002). In the case of harassment, social media facilitates individual- and network-level social interactions which serve to establish and reinforce symbolic boundaries, drawn by making moral distinctions between groups.

This emphasis on morality is crucial given psychologists William Brady and Molly Crockett’s MAD model of social contagion, which maintains that people are “*motivated* to share moral-emotional content based on their group identity; that such content is especially likely to capture *attention*; and that the *design* of social media platforms interacts with these psychological tendencies to further facilitate its spread” (Brady et al., 2020). People are morally outraged when they believe a moral norm has been violated, motivating them to shame and punish the violators (Crockett, 2017). While observing immoral actions is relatively infrequent in day-to-day life, it is constant on contemporary social media, where public shaming, “hot takes,” and clickbait are omnipresent.<sup>2</sup> Indeed, content that engages moral emotions is more likely to be shared within ideologically bounded groups (Brady et al., 2017), and because of this, may be promoted more frequently by recommendation algorithms (Crockett, 2017).

As a result, people are far more likely to encounter a moral norm violation online than offline.

In this context, “morality” refers to “ideas, objects or events typically construed in terms of the interests or good of a unit larger than the individual (e.g., society, culture, one’s social network)” (Brady et al., 2020). As a result, justifications for harassment are sometimes scaffolded through theories or reframing efforts that label someone or something immoral. For example, in my work with Robyn Caplan, we found that the term “misandry” is used by Men’s Rights Activists to construct an image of “feminism” as strategically harming and oppressing men and boys, thus justifying any attack on feminists as a move to stop misandry (Marwick & Caplan, 2018). Similarly, Dianna Anderson discusses the labeling of actor and writer Lena Dunham as a “child molester” based on a story in her autobiography that details how she touched her younger sister; this type of early childhood sexual exploration is entirely normal, but was used to paint the controversial Dunham as a bad person given universal condemnation of pedophiles (Anderson, 2018). In other cases, a dossier of misdeeds may be put together, known in internet slang as “receipts.” For example, in early 2020, writer Tracie Egan Morrissey posted a series of quotes from actress Jameela Jamil to Instagram, arguing that Jamil had inconsistently described her health concerns and past employment and concluding that Jamil had Munchausen syndrome. Jamil received waves of criticism as a result (Hampton, 2020).

In these cases, the moral violations named in the accusation position the accused as deviant because their behavior violates group norms. While Jamil publicly framed herself as a body-positive feminist activist, other feminists criticized her for taking up this cause as a conventionally attractive, thin actress (Dickson, 2018). In 2020, Jamil was hired to judge HBO’s voguing competition *Legendary*, a dance originating in queer Black and Latino ballroom culture, to which Jamil has no connection. After widespread online criticism, she came out as queer, thus establishing her legitimacy as a potential judge. Morrissey’s accusation built upon these critiques to position Jamil as an opportunist who failed to meet the strict norms of feminist activism. In this case, moral outrage established *symbolic boundaries* between authentic feminists, body-positive activists, and members of the LGBTQ+ community, and Jamil, who to her accusers was deceitfully taking advantage of such discourses for her own career advancement.

According to social identity theory, such symbolic boundary-making shores up the self-image of those criticizing the target (Riek et al., 2006). When people identify as part of a social group or category, they view people like themselves as the “in group” and people unlike themselves as the “out group” (Stets & Burke, 2000, p. 225). Because social groups provide shared identity and define one’s self-concept, people are deeply invested in the social status of their groups (Stephan et al., 2009). Extensive empirical studies have found that people favor members of their in-group; when they feel that their

social group is threatened, strong bias against the out-group emerges (Hogg, 2016). Thus, attacking an out-group responds to a perceived symbolic threat and reflects a desire to protect one's group status and thus self-image. In the United States, partisan and ideological identities are increasingly polarized and differentiated, begetting increased in-group identification and out-group denigration (Mason, 2018).

Because social norms around gender are both pervasive and persistent, women—especially women in the public eye who violate traditional norms of feminine quietude—experience disproportionate online harassment based on their gender (Citron, 2014; Sobieraj, 2020). Maass et al. apply social identity theory to sexual harassment, arguing that it may be “one strategy to protect or restore the male's threatened gender identity” (Maass et al., 2003 p. 854). Similarly, philosopher Kate Manne (2017) defines misogyny as the “hostile or adverse consequences” visited upon women who violate patriarchal norms (p. 13). Taking part in out-group denigration may feel pleasurable and give oneself a sense of moral righteousness, much as Manne (2017) describes the feeling of misogyny:

If it feels like anything at all, it will tend to be *righteous* . . . It often feels to those in its grip like a moral pursuit, not a witch hunt. And it may pursue its targets not in the spirit of hating women but rather, of loving justice. (p. 20)

To summarize, in morally motivated networked harassment, a member of a social network accuses a target of violating the network's norms, triggering moral outrage. Network members send harassing messages to the target, reinforcing their adherence to the norm and signaling network membership. Frequently, harassment results in the accused self-censoring and thus regulates speech on social media. To test this model, I draw from a corpus of harassing incidents gathered from interviews with people who identify as having experienced harassment.

## Method

This article draws from a corpus of semi-structured interviews with people who identify as having experienced online harassment ( $n=28$ ) and workers at Trust & Safety teams at various social media platforms ( $n=9$ ). The first group of participants were recruited using open calls posted on Twitter, Reddit, and Craigslist for people over 18 years who had experienced online harassment, defined in recruitment materials as

being called offensive names, having someone try to embarrass you on purpose, being physically threatened online, having sensitive personal information exposed or your privacy invaded, having rumors spread about you online, being sexually harassed or cyberstalked, or being harassed over a long period of time.

All interviews were conducted by the author and took place via phone or video chat (Zoom or Skype), depending on the participants' preference; as a result, some included audio and

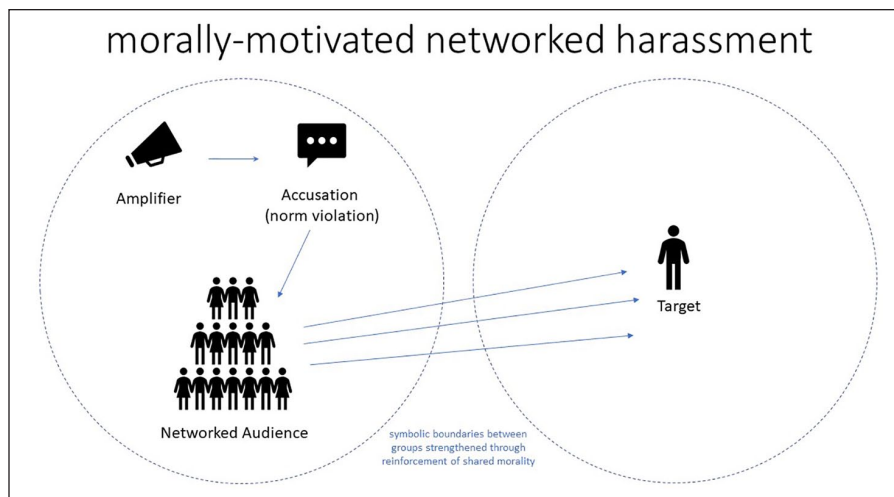
some audio and video. Interviews lasted between 30 and 90 min. Interviews followed a semi-structured protocol in which participants were asked about their experiences with online harassment, its effects, and their thoughts on online harassment in general (Appendix A). Subjects received a US\$20 incentive for participation, although a significant minority asked for it to be donated to a charity instead.

The group of Trust & Safety participants were recruited through email and LinkedIn, using pre-existing industry contacts to identify potential participants. Participants were asked about their company's procedures for dealing with unwanted user behavior and their experiences working in Trust & Safety (Appendix B). These subjects did not receive an incentive. These interviews were conducted via Zoom.

All interviews were audio-recorded and transcribed by an outside transcription company with a confidentiality agreement. Transcripts were imported into the qualitative data analysis software MaxQDA. I coded a subset of interviews ( $n=5$ ) to generate the initial codebook, which was developed iteratively through initial coding as coding continued (Corbin & Strauss, 2007). Initial coding allows the researcher to closely examine the data and identify potential categories (Saldana, 2009). Coding was done at both the *content* level (explicit statements by the interviewee) and *assumptions* and cultural discourses underlying the content (Gilligan & Brown, 1992). The second phase of coding involved focused coding to highlight significant and frequent codes (Charmaz, 2006).

The MMNH model was developed as interviews progressed, partly from interview data and partly from immersion in the literature. Later interviews explicitly asked participants about the model as a form of member checking, including all the interviews with Trust & Safety workers, who are exposed to hundreds or thousands of harassing incidents as part of their work (Creswell & Miller, 2000). The model was refined based on feedback from participants—for example, the concept of “attack vectors” came directly from member checks.

Participants who had experienced online harassment ranged in age from 18–49 years with an average age of 30.5 years. In all, 18 identified as women, 7 as male, and 3 as non-binary; 1 person identified as trans, 21 identified as White, 3 as Asian or Asian American, 1 as Black, and 2 as Middle Eastern; 13 identified as heterosexual, 1 as “mostly straight,” 8 as bisexual, 3 as queer, 1 as pansexual, 1 as lesbian, and 1 as gay. Participants were located primarily in the United States (17), although other participants lived in the United Kingdom (4), Europe (4), Canada (3), and Africa (1). All participants spoke English. Of the Trust & Safety participants ( $n=9$ ), 5 identified as women and 4 as men. Trust & Safety participants were not asked about their age, race, or sexuality to maintain pseudonymity given the small number of people working in the industry. All participant names and potentially identifiable information have been pseudonymized. In some cases, the specifics of the harassment or



**Figure 1.** Morally Motivated Networked Harassment.

platform features have been obscured to prevent possible identification. This study was approved by the University of North Carolina at Chapel Hill Institutional Review Board, IRB number 18-1916.

One might ask why I interviewed people who *experienced* online harassment when attempting to determine *motivations* for harassment. Extensive efforts to recruit people who identified as perpetrators of harassment, who had been accused of harassment, or even who admitted to having engaged in online conflict, were unsuccessful.<sup>3</sup> This may be because people do not want to admit that they engage in socially unacceptable behaviors like harassment, or because they considered their actions to be justified and therefore not harassment. However, participants provided valuable insight on the impact of harassment from the complete network of perpetrators (given that networked harassment involves dozens or hundreds of harassers), and from the perspectives of platform employees (who often speak to perpetrators while moderating harassment), as well as a diverse corpus of examples. However, there remains a need for research that includes the perspective of perpetrators.

### The MMNH Model

On social media multiple social contexts coexist (Marwick & boyd, 2011). This phenomenon of context collapse means that groups with diametrically opposed values—such as the far-left and the far-right, or the fat-positive and the fat-phobic—are visible to each other. Given this, harassment is a tactic that can be used by people across the political spectrum to reinforce norms. Some norm violations are widely held, such as hunting endangered animals, or letting a pet dog defecate on the subway (Klang & Madison, 2018). Others are specific to a community and do not map to political positions. For instance, one participant was harassed for writing critical essays about queer themes in a horror novel.

Although most research on harassment has focused on harassment by far-right or misogynistic actors (Burgess & Matamoros-Fernández, 2016; Citron, 2014), my participants included people who identified ideologically across the political spectrum, and as men, women, and nonbinary. Thus, the model based on these data is designed to apply to networked harassment regardless of the target or amplifier’s political or ideological leanings.

Networked harassment (Figure 1) typically begins by identifying one or more norm violations (the *accusation*) and tying it to a specific person, brand, or organization (the *target*), together creating a *justification* for harassment. This accusation is promoted by one or many key accounts or network nodes, such as highly followed social media accounts or influencers (the *amplifier*). Often, but not always, the *amplifier* is in a different social network than the *target*. Members of the amplifiers’ *networked audience*, who share an ideological or moral framework, individually send ad hominem attacks, insults, slurs, and in the worst cases, threats of death, rape, and violence to the accused (brigading, dogpiling, or “calling out”). Individual targets typically experience stress, depression, and other psychological harms, frequently resulting in self-censorship and withdrawal from social media participation. Simultaneously, the ideological consensus of the accusing network is reinforced through a common enemy and the symbolic boundaries between contexts are reinforced. Thus, harassment becomes a regulating force in which speech is removed from the public sphere.

### The Accusation

The accusation describes a violation of a social norm that serves to justify harassment from the networked audience. The accusation could be a single incident, or a body of evidence built up over time. Individuals in my sample were

harassed for a wide variety of incidents such as tweeting, “Man, I hate white men” in response to a video of police brutality; criticizing the pre-Raphaelite art movement; refusing to testify in favor of a student accused of sexual harassment; banning someone from a popular internet forum; criticizing expatriate men for dating local women; and celebrating the legalization of gay marriage in the United Kingdom by posting gifs of the television show “Sherlock.” While some of these accusations may not appear to violate moral norms, delving into specific cases shows how critique, disagreement, or just plain dislike of the target are reframed by the amplifier or their networked audience to accuse the target of moral violations.

**Attack Vectors.** Despite framing harassment as a tactic used across ideological spectra, it must also be linked to structural systems of misogyny, racism, homophobia, and transphobia, which determine the primary standards and norms by which people speaking in public are judged. For example, while a nonbinary individual might be harassed over comments about anything, their nonbinary status will frequently become an “attack vector,” information security slang for an exploitable system vulnerability. In other words, they will often be attacked for being nonbinary even if it has nothing to do with the matter at hand. Khalid (23, Middle Eastern), who identifies as non-binary and Muslim, explained how harassment breaks down along intersectional lines:

In my bio, I include my sexuality, my gender identity, but I also include my religion and what I noticed is, generally when . . . My bio changes throughout and when I include my religion in that, there’s more of a chance of people catching onto that detail and harassing me on that basis, rather than just “That’s some guy on the Internet, we really don’t care.”

Khalid finds that when they explicitly define as Muslim, they are bombarded with Islamophobic statements. Similarly, many of the women I interviewed experienced gender-specific forms of violence such as rape threats, pornographic imagery, gendered slurs, sexually explicit threats, and so forth. While my sample included many men, their gender status *as* men was not an attack vector. In other words, they were not subject to harassment based on their gender. Such attack vectors generally focus on marked characteristics: women marked when men are unmarked, people of color marked whereas Whiteness goes unmarked, and so forth (Brekhus, 1998).

However, examples of harassment from left-wing networks (broadly defined) included some attacks based on Whiteness or maleness. Heath (42), who is biracial but White-passing, experienced harassment based on his association with law enforcement. He told me,

people would follow or send a message on [Twitter] or [Reddit], and the message would be, “You need to quit your job. You’re

harming people. You’re nothing but a white man who’s oppressing,” which is hard enough, but it was the, “You should kill yourself. You should die.” stuff [that was really difficult].

Heath considers himself a progressive and recognized that such harassment is rooted in a complex history of police brutality against people of color, but still resented the accusations of White supremacy leveraged against him. Such experiences suggest that attack vectors originating in those aligned with historically marginalized groups may focus more on unmarked characteristics, and points to the need for more research on left-wing harassment.

### Justification

Accusations of immorality according to the moral norms of the networked audience serve as justifications for networked harassment. Given the serious consequences that harassment has on the target, the accusation must go beyond simple dislike or bias; it must be strongly rooted in shared values or norms. As a result, justifications often seem to be constructed retrospectively. For example, when Anita Sarkeesian first launched her *Feminist Frequency* web series, her attackers accused her of advocating censorship. As the harassment continued, perhaps when it became clear that Sarkeesian was not promoting censorship, the accusations morphed. Sarkeesian describes her detractors’ image of her as a “folk demon” or “Disney villain,” based on an “information cascade” (XOXO Festival, 2014). This cascade consists of a dossier of primarily false information that makes her out to be duplicitous, ignorant about video games, dismissive of female gamers, and even violent. It includes fabricated pictures, reports, interviews, and social media posts, such as a tweet of a picture of Gucci pumps with the caption “Buying 1,000 dollar shoes.” She says,

Information cascades occur when people rapidly repeat or share information from others without first verifying its validity, so falsehoods about me are initially pushed by detractors who use them to spam 4chan and Reddit as a way of provoking rage and rallying more people to join the crusade against me. As the disinformation spreads through social media, it takes on a life of its own. It’s bouncing from Twitter to Facebook to Tumblr to YouTube and back again. Once the cascade reaches a critical mass, it no longer matters what the facts are. (XOXO Festival, 2014)

The networked nature of social media allows accusations or “receipts” to spread rapidly without fact-checking.

Similarly, Men’s Rights Activists coined *misandry* to reframe feminism as a movement actively trying to hurt men and boys, thus justifying attacks on feminists (Marwick & Caplan, 2018). In an analysis of YouTube “response videos” which react to and debate other videos, Lewis et al. (2021) found that creators used these “debates” to paint creators as immoral or offensive. For example, a video critiquing the

American mythology of Thanksgiving as colonialist and racist was reframed as an anti-White video advocating for a “ban” on Thanksgiving, a much less sympathetic prospect (Lewis et al., 2021). Participants in this study described similar reframing.

Nicole (White, 30) is a body liberation advocate, a proud fat woman who’s written two books and given a TED talk about releasing oneself from body hatred and diet culture. Always a target for online trolls and abuse, her experiences worsened once a “fat hate” blog wrote about her Patreon, a crowd-funding site for creators to solicit financial support from audiences. Nicole asked her readers to donate money so she could write full-time and work on a body positivity conference. This triggered a flood of networked harassment and online abuse. She received hundreds of messages a day on every online platform, calling her fat and disgusting. She was very worried by a post on a “fat hate” subreddit that described running into her at neighborhood Trader Joe’s, encouraging other people in the area to look for her.<sup>4</sup>

The messages were what you might expect—heavily sexist and fatphobic—but focused on the idea that she was making money illegitimately through Patreon. She told me:

The Internet couldn’t regulate [my work], although they tried. They tried to calculate hours. They tried to calculate what I was eating. They tried to look at my conference board picture and the tables and see how much money on my Patreon we were spending on coffee and like all of these really weird things. But they couldn’t really regulate how much work that I was doing for the conference.

To members of the subreddit, what Nicole was doing was wrong. They believed she was profiting financially from promoting unhealthy behavior, and potentially scamming people out of their hard-earned money. The subreddit collectively constructed a *moral justification* for its harassment of Nicole. The harassment affected Nicole deeply. She has mostly stopped writing publicly and instead focuses on her book projects. Nicole strongly believes that being a happy, successful fat woman—especially one with the hubris to ask for recompense for her creative efforts—makes her a target.

### Amplification

In the process of *amplification*, an accusation spreads quickly through a network, often because it is signal-boosted by a popular account or community. For example, Adrienne (White, 32) was harassed for a tweet about gender diversity in young adult novels:

There was one main person who is famous in our community for directing harassment at people who has a pretty big following, like 50,000 followers or something like that. She was the one who said that this stuff [that I posted] was super racist. She pointed people at me, and her followers are famous for harassing people, anyone who she disagrees with publicly.

In this case, Adrienne’s tweet got little traction until it was picked up and amplified by the highly followed account. Similarly, Constance (White, 33) experienced harassment within a fan community on Tumblr. She received death threats after disagreeing with a “big name fan’s” interpretation of a novel:

This one person who was a fairly . . . big name person, who I *don’t* think is the one who was sending me death threats, but who I think contributed to an atmosphere where people felt I was fair game? She was in the habit of having people ask her authoritative questions, and just arbitrating them? . . . Finally, I wrote a response to one of her posts . . . It turned into a whole thing because she was just *deeply* offended that anyone would say that she could be wrong about something.

Because amplifiers (highly followed nodes, in network terms) have so many viewers, they are able, consciously or not, to direct harassing behavior. It is often not the amplifiers themselves but those who follow them who engage in such behavior, making it difficult to ascertain responsibility (Lewis et al., 2021).

Elise, the Asian American musician we met earlier, became concerned about the impact of her tweet criticizing a restaurateur when it was retweeted by “some really big name accounts, especially accounts that have really large followings, like tens of thousands or hundreds of thousands of followers.” Because the followers of these “big name” accounts did not know Elise or the context of her tweet, Elise worried that they would dogpile on the restaurateur because “[they] are ready to attack anything retweeted by these accounts because they tend to focus on like social justice. So I was like, ‘This could be bad.’” In this case, Elise’s critical tweet was amplified by progressive accounts, possibly causing harassment for the target of her critique. However, when it was amplified by “anti-SJW”<sup>5</sup> accounts, she experienced harassment. Thus, amplification goes both ways.

### The Networked Audience

The “networked audience” is the “real and potential viewers for digital content that exist within a larger social graph. These viewers are connected not only to the user, but to each other, creating an active, communicative network” (Marwick & boyd, 2011, p. 16). Distinguished from the presumably faceless mass of the broadcast audience, the networked audience is connected in a web of complex social or symbolic relationships, and often shares cultural commonalities. If we consider ideological polarization, different networked audiences with very different moral standards and values may be co-present on social media (Mason & Wronski, 2018). They can be the audience for a particular account or person who acts as an amplifier, or they can be members of an online community that is amplifying the accusation, as in Nicole’s case.

The presence of the networked audience is significant for several reasons. First, content that contains moral outrage spreads more rapidly within like-minded networks than other content on social media (Brady et al., 2017). Second, certain forms of internet conflict, known as “drama,” are more common in front of audience members, who may see them as forms of entertainment (Miller, 2016; Regan & Sweet, 2015). Third, the audience participates directly in networked harassment by responding to the amplified justification. The thousands of tweets or negative comments received by my participants, in most cases, were not from high-profile accounts, but members of their audience or of online communities:

When someone has 25K Twitter followers, they pile on really quickly, and it sort of becomes, especially in this case where it’s very conspiratorially-minded thinking, that the accusations and the allegations sort of start to compound and build up on each other. (Keith, 35)

I tweeted about [a screenwriter] one time because he’s this guy who keeps getting major blockbuster movie deals when his movies pretty much flop universally, and he name searched his name on Twitter, because I didn’t tag him. He came and found it and then retweeted it to his followers and said something derogatory about me and then at that time he had a million followers who are all these angry nerd boys and he didn’t get suspended, that was like directing targeted harassment at me, you know? (Danielle, 27)

Keith was accused of being a Russian disinformation theorist by a highly followed conspiracy theorist on Twitter. The networked audience of his accuser worked together to create a complex justification for his harassment: namely, that he was an enemy operative who had insider knowledge from his Russian handlers about the NSA (in reality, Keith is a graduate student with a Slavic last name who is interested in Russian history, not an intelligence professional). Danielle’s harassers followed a popular screenwriter. Her characterization of his followers as “angry nerd boys” describes a networked audience with a shared set of values.

### The Harassment

Networked harassment itself can take a variety of forms but must be coordinated at scale. A single harassing message is not networked harassment; neither are dozens of messages coming from a single person (described elsewhere as *dyadic harassment*). Networked harassment involves many individuals sending messages, emails, or phone calls within a relatively short amount of time. In my sample, participants described such harassment:

So many people came into my mentions and were yelling racial slurs at me even though I’m white, and so many different things like “You’re just a stupid bitch. You don’t know what you’re

talking about. This is America. You should be proud of this. You just need to get back in the kitchen instead of giving your opinion,” and a bunch of ridiculous stuff like that. (Ava)

One time I found a Reddit thread that was about bashing me and somebody said I should be forcibly sterilized, like just random little pockets of internet hatred. I’ve actually, I’ve seen a lot of stuff on Reddit . . . when I was writing about entertainment and stuff I’ve seen a lot of just mini flurries of like I get linked to in some Reddit forum and then I’m like okay, where are all these people in my Twitter mentions coming from, oh it’s coming from this Reddit forum. (Danielle)

The content of these messages may include accusations of moral violation, hateful speech, profanities, insults, death, and rape threats, publicizing private information or photographs, dossiers or receipts of misdeeds, gory or pornographic imagery, and the like, all of which appeared in my sample. As might be expected, such harassment leads to significant negative outcomes for the target.

### Outcomes

According to the MMNH model, networked harassment has three primary outcomes. First, echoing previous research, targets of harassment experience depression, anxiety, and other negative emotional consequences. Second, these emotional consequences often lead to self-censorship on the part of the target, causing them to decrease their online public presence. Finally, networked harassment reinforces the norms of participating networks, solidifying the boundaries between them and others.

Previous research indicates that online harassment has significant negative impacts on people’s lives (Blackwell et al., 2017; Sobieraj, 2020), including emotional and psychological difficulties (Duggan, 2017). This finding was echoed by most participants. Elise, for example, described her reaction as “seriously depressed.” She said,

I was depressed that week. It was just not a good time at all. It took me, like, another week afterwards to kind of go back to normal, to be like, “I feel okay leaving the house again, I feel okay checking email again. I feel okay to post stuff on Twitter.”

Interestingly, my interviewees frequently preceded discussions of negative emotional consequences with a feigned dismissal that they “knew” online harassment should not bother them:

It made me feel really awful. I know reasonably that it’s just a random person hiding behind their keyboard and projecting getting his feelings out on whatever, but it was just hurtful that they chose me and chose to use those words. (Ava)

I’m sure by the standards of harassment [the incidents were] very mild, but they were both very traumatic to me. (Adrienne)



It really upset me, and it really bummed me out. I felt like maybe the only people I could turn to were my wife and my close friends who were also targeted. It didn't really feel like anybody understood or really had much sympathy for what was going on in that moment, even though it was very similar to other things, and that, not having anybody to talk to about it, especially someone above me or someone in a position of power to help me get through that moment, that kind of made me feel very alone in this situation. (Keith)

As the interviews went on, I found myself assuring participants that almost everyone I talked to had suffered mentally from harassment, even if it was “just a random person hiding behind their keyboard.” Participants described depression, isolation, anxiety, and stress, among other negative consequences, and often seemed to feel guilty or ashamed that they felt bad. This points to a possible intervention: emotional support for those going through harassment must include messaging that it is normal to feel bad when attacked online. This also suggests that trying to diminish people's experiences by shrugging them off as something that only happens online is counterproductive.

As we have seen in the case of fat activist Nicole, such emotional fallout frequently led participants to pull back on their online participation. Elise said,

I still feel like I'm still not completely myself on Twitter because now I'm more careful about what topics I talk about. I don't really talk about racism and sexism so much anymore, or I don't really talk about politics, I am a lot more cautious now.

By choosing not to participate anti-racist activism in the digital public sphere to avoid any future harassment, Elise changed her behavior to conform with the norms of the networked audience who attacked her. This also concurs with previous research. One study found that more than half of women 15–29 years censor themselves online to avoid harassment (Veletsianos et al., 2018). Other scholarship has found that female scholars, journalists, and politicians self-censor themselves online due to both online harassment and the *fear* of online abuse (Binns, 2017; Chen et al., 2020; Sobieraj et al., 2020), culminating in what Carter Olson and LaPoe (2018) call a “digital Spiral of Silence.”

Thus, self-censorship is another way in which networked harassment functions as a norm reinforcement technique. The networked audience engaged in harassment successfully reinforces its moral norm by discouraging the target from violating it in public. And, because networked harassment is more likely to happen to women, trans and non-binary people, sexual minorities, and people of color, networked harassment causes minoritized voices to be systematically eliminated from the public sphere. Networked harassment also reinforces symbolic boundaries between groups by creating moral distinctions between *us* and *them*. As discussed, polarization and partisanship involve both in-group solidarity and out-group animus. Although most research on harassment has focused

on harassment by right-wing and conservative networks, it is a tactic used by groups across the ideological and political spectrum.

Specifically, in left-leaning communities, “callout culture” functions in much the same way as harassment does, to punish individuals for moral transgressions typically involving racism, sexism, ableism, transphobia, and so forth (Clark, 2020). This is, perhaps, a more sympathetic proposition. For instance, more than one participant noted that they believed harassment of neo-Nazis or racists was justified. As Danielle, a White woman who identified as a feminist, put it:

There's the case of where people could march in white supremacist rallies, there are whole Twitter accounts dedicated to “this is this person, does anyone recognize them, can we find out where he works.” In that case I do think they deserve to be outed, I guess. I don't think anyone deserves a death threat or to have where they live posted online, but I think there are things that are so toxically extreme that the only thing that pushes it back is public outcry . . . Should people who march in white supremacist rallies be afraid of their job finding out? I think yeah.

In other cases, even people who had experienced harassment were sympathetic to the political positions of the perpetrators and believed that it was not the tactic that was misguided, but the target. This suggests that if the moral justification for harassment resonates with the morality of the individual, they may believe the harassment is justified even though they have experienced the negative consequences of harassment themselves. These individuals reinforced their own moral justification for harassment even while acknowledging the harm that harassment did to them and others they knew who had experienced it.

## Limitations

This model is based on the considerable literature on online harassment, as well as a diverse corpus of harassing incidents and experiences. However, there are two clear gaps in the literature. First, more research is needed on harassment perpetrated by left-leaning groups and individuals, particularly about attack vectors, as previously noted. Second, this study did not interview people who have harassed others; such a project could shore up the justification aspect of the model. Finally, future research could test the MMNH model against a larger and more diverse corpus of harassing incidents to determine whether there are incidents of networked harassment that are not morally motivated.

## Conclusion and Implications

The MMNH model has several implications for technology companies and scholars. For researchers, this model of harassment can be applied to a vast array of networks beyond the American right and left-wing. Highly partisan, polarized

sets of public opinions along the ideological spectrum coexist on social media. Networked harassment is a tactic used across political and ideological groups and, as we have seen, by groups that do not map easily to political positions, such as conflicts within fandom or arguments over business. More research is needed on different types of harassment to further refine this model.

Second, the MMNH model suggests that moderation that examines individual content pieces to determine whether they violate community standards may miss the forest for the trees: that is, the amplifier effect of a major network node shaming an individual, resulting in networked harassment. The Terms of Service of most websites adhere to United States laws around harassment, which presume the dyadic model of one individual repeatedly harassing another. This means that identifying networked harassment using individual-level models of harassment is very difficult. As Adrienne said,

The Twitter staff might recognize it as harassment if one person sent you 50 messages telling you suck, but it's always the one person [who] posted one quote tweet saying that you suck, and then it's 50 of their followers who all independently sent you those messages, right?

Legal definitions of harassment presume ongoing harassment by the same person rather than a network. As Justine (Korean American, 25) explained,

And then when I started experiencing [harassment] online, was also like, the fact that I don't know who this person is in many ways kind of universalizes it, so I feel like anybody that I walk next to in the street, like could be the harasser, but in terms of appealing to, like, a legal or policy definition, it doesn't fit, because this one person shared, like, a selfie of me on the school gossip website, but there hasn't been a second offense, and it's unclear if it was a joke, or not.

Social platforms must recognize the amplifying effect of highly followed networked nodes accusing other users of nefarious deeds. For better or for worse, those with a larger audience bear a greater responsibility for their online actions if they silence others. Given that the consequences of networked harassment can be the systemic suppression of minoritized voices, it is crucial that platforms work diligently to prevent the chilling effects of harassment.

Another possible area for intervention is reminding users of platform norms that prohibit harassment. J. Nathan Matias' work with Reddit has shown that displaying community rules that reflect subreddit norms (removing abusive or off-topic comments, for example) made newcomer comments less likely to be removed, and increased participation rates (Matias, 2019). This would not stop behavior in communities where harassment is condoned, such as the Fat People Hate subreddit that harassed Nicole, but it suggests a possible source of friction for platforms like Twitter.

Ultimately, conceptualizing harassment as *morally motivated* and understanding it as a technique of *norm reinforcement* explains why people participate in it, a necessary step to decreasing it. This model may open creative solutions to harassment and content moderation. MMNH also recognizes that harassment, while more endemic to minoritized communities, may be experienced by people from a wide variety of identities and political commitments, suggesting many possibilities for future research. Current technical and legal models of harassment do not protect against networked harassment; by providing a new model, I hope to contribute to lessening its prevalence.

### Acknowledgements

I am indebted to Molly Crockett, Lindsay Blackwell, Kat Lo, Nate Matias, Robyn Caplan, Danielle Citron, danah boyd, Amy Bruckman, and Adrienne Massanari for inspiration, advice, guidance, and foundational work without which this paper would not have come to pass. Thank you to Robyn Caplan, danah boyd, and Kat Lo for recruiting assistance. Thank you to Shannon McGregor and Katie Furl for reviewing the final draft of the paper and providing useful feedback. Thank you to the Department of Communication, the Institute for Arts and Humanities at UNC Chapel Hill, and the Data & Society Research Institute for funding the research sabbaticals during which this fieldwork was conducted. Thank you also to the attendees of the MIT Workshop on Harassment, the Yale Symposium on Platform Governance, the NetGain Dis/Mis-information, Dangerous Speech, and Democracy Meeting, and the Fracturing Democracy: The Erosion of Civil Society in a Shifting Communication Ecology Symposium at the University of Wisconsin–Madison for generative discussions. Finally, and most importantly, thank you to my participants for their willingness to share their stories with me.

### Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.


### Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was funded by the Institute for the Arts and Humanities at UNC Chapel Hill and the Data & Society Research Institute.

### Ethical Approval

This study was approved by the University of North Carolina at Chapel Hill Institutional Review Board, IRB Number 18-1916.

### ORCID iD

Alice E. Marwick  <https://orcid.org/0000-0002-0837-6999>

### Notes

1. While Elise does not know who amplified her tweet, she believes it was a conservative account given that most of the profiles of people who harassed her included content that marked them as American conservatives (e.g., support for President Trump and MAGA).

2. It is relatively rare to be an eyewitness to immoral actions such as abuse, murder, theft, and so forth. However, more common but milder immoral actions such as “cutting in line” are often met with intense wrath.
3. The recruitment materials noted that we were seeking people who had experienced online harassment, people who had been accused of online harassment (“being banned from a community, added to a blockbot, reported for abuse on a social media site, and so forth”), and people who had engaged in online conflict (“calling someone offensive names, exposing someone’s sensitive personal information [doxing], encouraging others to bother someone [brigading], spreading online rumors about another person, deliberately invading another person’s privacy, threatening someone online, and so forth.” The latter categories did not yield any respondents.
4. A “subreddit” is an online forum on Reddit dedicated to a particular topic. There are thousands of subreddits focusing on everything from fandom to knitting to weight loss to memes.
5. “Social Justice Warrior,” a pejorative term used by the right-wing for progressive (feminist, anti-racist) activism.

## References

- Anderson, D. E. (2018). *Problematic: How toxic callout culture is destroying feminism*. University of Nebraska Press.
- Banet-Weiser, S., & Miltner, K. M. (2016). #MasculinitySoFragile: Culture, structure, and networked misogyny. *Feminist Media Studies*, 16(1), 171–174.
- Barton, A., & Storm, H. (2014). *Violence and harassment against women in the news media: A global picture*. International Women’s Media Foundation. <http://www.iwmf.org/our-research/journalist-safety/violence-and-harassment-against-women-in-the-news-media-a-global-picture/>
- Billingham, P., & Parr, T. (2020). Enforcing social norms: The morality of public shaming. *European Journal of Philosophy*, 28(4), 997–1016.
- Binns, A. (2017). Fair game? Journalists’ experiences of online abuse. *Journal of Applied Journalism & Media Studies*, 6(2), 183–206. [https://doi.org/10.1386/ajms.6.2.183\\_1](https://doi.org/10.1386/ajms.6.2.183_1)
- Blackwell, L., Chen, T., Schoenebeck, S., & Lampe, C. (2018). When online harassment is perceived as justified. In *Proceedings of the 12th International Conference on Web and Social Media (ICWSM 2018)* (pp. 22–31). AAAI. <https://www.aaai.org/ocs/index.php/icwsm/icwsm18/paper/view/17902>
- Blackwell, L., Dimond, J., Schoenebeck, S., & Lampe, C. (2017). Classification and its consequences for online harassment: Design insights from heartmob. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW), 1–19.
- Brady, W. J., Crockett, M., & Van Bavel, J. J. (2020). The MAD Model of Moral Contagion: The role of motivation, attention and design in the spread of moralized content online. *Perspectives on Psychological Science*, 15(4), 978–1010. <https://doi.org/10.1177/1745691620917336>
- Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 114(28), 7313–7318. <https://doi.org/10.1073/pnas.1618923114>
- Brekhus, W. (1998). A sociology of the unmarked: Redirecting our focus. *Sociological Theory*, 16(1), 34–51.
- Burgess, J., & Matamoros-Fernández, A. (2016). Mapping socio-cultural controversies across digital media platforms: One week of #gamergate on Twitter, YouTube, and Tumblr. *Communication Research and Practice*, 2(1), 79–96.
- Carter Olson, C., & LaPoe, V. (2018). Combating the digital spiral of silence: Academic activists versus social media trolls. In J. R. Vickery & T. Everbach (Eds.), *Mediating misogyny: Gender, technology, and harassment* (pp. 271–291). Springer International Publishing. [https://doi.org/10.1007/978-3-319-72917-6\\_14](https://doi.org/10.1007/978-3-319-72917-6_14)
- Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis* (1st ed.). SAGE.
- Chen, G. M., Pain, P., Chen, V. Y., Mekelburg, M., Springer, N., & Troger, F. (2020). “You really have to have a thick skin”: A cross-cultural perspective on how online harassment influences female journalists. *Journalism*, 21, 877–895.
- Citron, D. (2014). *Hate crimes in cyberspace*. Harvard University Press.
- Clark, M. (2020). DRAG THEM: A brief etymology of so-called “cancel culture.” *Communication and the Public*, 5(3–4), 88–92. <https://doi.org/10.1177/2057047320961562>
- Corbin, J., & Strauss, A. (2007). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (3rd ed.). SAGE.
- Creswell, J. W., & Miller, D. L. (2000). Determining validity in qualitative inquiry. *Theory into Practice*, 39(3), 124–130.
- Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour*, 1(11), 769.
- Dickson, E. J. (2018, December 4). *How Jameela Jamil built a brand around body positivity*. <https://www.vox.com/the-goods/2018/12/4/18124392/jameela-jamil-good-place-body-positivity>
- Duggan, M. (2017). *Online harassment 2017*. Pew Research Center. <https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/>
- Eikren, E., & Ingram-Waters, M. (2016). Dismantling “you get what you deserve”: Toward a feminist sociology of revenge porn. *Ada: A Journal of Gender, New Media, and Technology*, 10. <http://adanewmedia.org/2016/10/issue10-eikren-ingramwaters/>
- Finkelhor, D., Mitchell, K. J., & Wolak, J. (2000). *Online victimization: A report on the nation’s youth*. National Center for Missing and Exploited Children.
- Gilligan, C., & Brown, L. M. (1992). *Meeting at the crossroads: Women’s psychology and girls’ development*. Harvard University Press.
- Hampton, R. (2020, February 22). Making sense of the dizzyingly complicated Jameela Jamil controversy. *Slate Magazine*. <https://slate.com/culture/2020/02/jameela-jamil-munchausen-accusations-bees-ehlers-danlos-explained.html>
- Henry, N., & Powell, A. (2015). Embodied harms: Gender, shame, and technology-facilitated sexual violence. *Violence against Women*, 21(6), 758–779. <https://doi.org/10.1177/1077801215576581>
- Hogg, M. A. (2016). Social identity theory. In S. McKeown, R. Haji, & N. Ferguson (Eds.), *Understanding peace and conflict through social identity theory: Contemporary global perspectives* (pp. 3–17). Springer International Publishing. [https://doi.org/10.1007/978-3-319-29869-6\\_1](https://doi.org/10.1007/978-3-319-29869-6_1)
- Jane, E. A. (2014). “Your a ugly, whorish, slut” Understanding E-bile. *Feminist Media Studies*, 14(4), 531–546.
- Jhaver, S., Chan, L., & Bruckman, A. (2018). The view from the other side: The border between controversial speech and

- harassment on Kotaku in Action. *First Monday*, 23(2). <http://firstmonday.org/ojs/index.php/fm/article/view/8232/6644>
- Klang, M., & Madison, N. (2018). Vigilantism or outrage: An exploration of policing social norms through social media. In B. Vanacker & D. Heider (Eds.), *Ethics for a Digital Age* (Vol. II) (pp. 151–165). Peter Lang.
- Klonick, K. (2015). A new taxonomy for online harms. *Boston University Law Review Annex*, 95, 53–55.
- Krook, M. L. (2017). Violence against women in politics. *Journal of Democracy*, 28(1), 74–88.
- Lamont, M. (2017). *Prisms of Inequality: Moral Boundaries, Exclusion, and Academic Evaluation*. Praemium Erasmianum Foundation.
- Lamont, M., & Molnár, V. (2002). The study of boundaries in the social sciences. *Annual Review of Sociology*, 28(1), 167–195.
- Lenhart, A., Ybarra, M. L., Zickuhr, K., & Price-Feeney, M. (2016). *Online harassment, digital abuse, and cyberstalking in America*. Data & Society Research Institute.
- Lewis, R., Marwick, A., & Partin, W. (2021). “We dissect stupidity and respond to it”: Response videos and networked harassment on YouTube. *American Behavioral Scientist*, 65, 735–756.
- Maass, A., Cadinu, M., Guarnieri, G., & Grasselli, A. (2003). Sexual harassment under social identity threat: The computer harassment paradigm. *Journal of Personality and Social Psychology*, 85(5), 853–870.
- Manne, K. (2017). *Down girl: The logic of misogyny*. Oxford University Press.
- Mantilla, K. (2013). Gendertrolling: Misogyny adapts to new media. *Feminist Studies*, 39(2), 563–570. <https://doi.org/10.2307/23719068>
- Marwick, A. (in press). *Finding gender in the network*. In *The Private is Political*. Yale University Press.
- Marwick, A., & boyd, d. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, 13(1), 114–133.
- Marwick, A., & Caplan, R. (2018). Drinking male tears: Language, the manosphere, and networked harassment. *Feminist Media Studies*, 18(4), 543–559.
- Mason, L. (2018). Ideologues without issues: The polarizing consequences of ideological identities. *Public Opinion Quarterly*, 82(S1), 866–887. <https://doi.org/10.1093/poq/nfy005>
- Mason, L., & Wronski, J. (2018). One tribe to bind them all: How our social group attachments strengthen partisanship. *Political Psychology*, 39(S1), 257–277. <https://doi.org/10.1111/pops.12485>
- Matias, J. N. (2019). Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences*, 116(20), 9785–9789. <https://doi.org/10.1073/pnas.1813486116>
- McGlynn, C., Rackley, E., & Houghton, R. (2017). Beyond ‘Revenge Porn’: The continuum of image-based sexual abuse. *Feminist Legal Studies*, 25(1), 25–46. <https://doi.org/10.1007/s10691-017-9343-2>
- Miller, S. A. (2016). “How you bully a girl”: Sexual drama and the negotiation of gendered sexuality in high school. *Gender & Society*, 30(5), 721–744. <https://doi.org/10.1177/0891243216664723>
- Nussbaum, M. C. (2009). *Hiding from humanity*. Princeton University Press.
- Regan, P. M., & Sweet, D. L. (2015). Girls and online drama: Aggression, surveillance, or entertainment. In J. Bailey & V. Steeves (Eds.), *Egirls, ecitizens: Putting technology, theory and policy into dialogue with girls’ and young women’s voices* (pp. 175–198). University of Ottawa Press.
- Riek, B. M., Mania, E. W., & Gaertner, S. L. (2006). Intergroup threat and outgroup attitudes: A meta-analytic review. *Personality and Social Psychology Review*, 10(4), 336–353.
- Saldana, J. (2009). *The coding manual for qualitative researchers*. SAGE.
- Sobieraj, S. (2020). *Credible threat: Attacks against women online and the future of democracy*. Oxford University Press.
- Sobieraj, S., Masullo, G. M., Cohen, P. N., Gillespie, T., & Jackson, S. J. (2020). Politicians, social media, and digital publics: Old rights, new terrain. *American Behavioral Scientist*, 64(11), 1646–1669. <https://doi.org/10.1177/0002764220945357>
- Stephan, W. G., Ybarra, O., & Rios Morrison, K. (2009). Intergroup threat theory. In T. D. Nelson (Ed.), *The handbook of prejudice, stereotyping and discrimination* (pp. 43–59). Psychology Press.
- Stets, J. E., & Burke, P. J. (2000). Identity theory and social identity theory. *Social Psychology Quarterly*, 63(3), 224–237. <https://doi.org/10.2307/2695870>
- Tokunaga, R. S. (2010). Following you home from school: A critical review and synthesis of research on cyberbullying victimization. *Computers in Human Behavior*, 26(3), 277–287.
- Veletsianos, G., Houlden, S., Hodson, J., & Gosse, C. (2018). Women scholars’ experiences with online harassment and abuse: Self-protection, resistance, acceptance, and self-blame. *New Media & Society*, 20, 4689–4708.
- Vitak, J., Chadha, K., Steiner, L., & Ashktorab, Z. (2017, February). *Identifying women’s experiences with and strategies for mitigating negative effects of online harassment*. Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (pp. 1231–1245). ACM. <https://dl.acm.org/doi/10.1145/2998181.2998337>
- XOXOFestival. (2014). *Anita Sarkeesian, Feminist Frequency—XOXO Festival (2014)*. <https://www.youtube.com/watch?v=ah8mhDW6Shs&t=35s>
- Ybarra, M. L., & Mitchell, K. J. (2008). How risky are social networking sites? A comparison of places online where youth sexual solicitation and harassment occurs. *Pediatrics*, 121(2), e350–e357.

## Author Biography

Alice E. Marwick is an associate professor of Communication and principal researcher at the Center for Information, Technology, & Public Life at the University of North Carolina. She studies the social, cultural, and political impacts of social media.

## Appendix A

### Interview Questions for People Who Have Experienced Harassment

- Tell me a little bit about yourself.
- What do you do in your free time? What are you passionate about?
- What social media do you use? What do you use the most?
  - Who do you think of as the audience for each social media platform?
  - Do you use the same username on each?

- You're part of this study because you've experienced online harassment. I'm sorry about your negative experiences online, but I'd like to ask you some questions about them so we can learn more about online harassment and potentially help other people who've experienced it. Please feel free to skip any question you don't feel comfortable answering.
  - Can you tell me what happened? How did you feel?
  - Do you know who harassed you? What do you think caused them to engage in this behavior?
  - Do you think your identity played a part in your harassment?
  - What helped you and what type of resources would have been helpful during this time? (mental health, police, support, financial, technological, etc.)
  - Did you talk about what was happening with friends online or offline?
  - What effects has the harassment had on you? Have you changed your internet behavior as a result?
- How do you protect yourself from harassment and abuse online?
- What do you think constitutes "harassment"?
- What do you think tech companies should do to reduce harassment? What should lawmakers do? Are there social situations?
- Do you think there are situations in which any of the behaviors we talked about are justified?
  - For example, a hunter posted a picture online of a lion he shot on Twitter and got tons of mean tweets and hate mail. Was this justified? Why?
  - A public relations executive made a joke about AIDS and Africa on Twitter and was fired from her job. Do you think that's justified?
  - Two guys at a tech conference joking around about "dongles" were fired after a picture of them was posted on Twitter. Do you think that's justified?
- Who do you think is most likely to get accused of harassment? Who is most likely to be harassed? Who is most likely to harass others?
- Why do you think people harass others?
- Say an online newspaper posts a story that women are harassed more than men online. Do you think this is

fair? How would you like to see the media discuss online harassment?

- Have you ever been accused of harassment?

## Appendix B

### Interview Questions for Trust & Safety Workers

- Can you walk me through the most common kinds of *unwanted user behavior* on your platform?
- How would you define "unwanted user behavior"?
- What attack vectors exist on your product?
- How is harassment defined internally?
- What actions do you consider harassing behavior? Can you give me examples?
- I've found that users define harassment in a wide variety of ways. How do you handle that?
- How do you escalate harassment?
- How are users "punished"?
- What is the profile of the user who engages in harassing behavior? Is there one?
- What types of people or networks are more likely to perpetrate harassing behavior?
- In your experiences, are there patterns in the types of people who are more likely to experience harassing behavior?
- What are the biggest challenges in combating harassing behavior?
  - How do you deal with context?
  - How do you deal with networked harassment?
- My research has led me to form a model of networked harassment as *morally motivated*. (Explain model). What do you think of this model? What is it missing?
- Do you feel that it's easy to advocate for T&S within a larger tech organization?
- How do you acquire domain knowledge about T&S?
- Do you talk to T&S people at other companies?
- Where do you think T&S as a field is today compared to where it was five years ago?
- Do you think there is a gendered component to harassment?
  - What about race?
  - Sexuality?
- What technological solutions do you see to combating harassing behavior?
  - Are there social solutions?
  - Legal/policy solutions?