When Does it Become Harassment?: An Investigation of Online Criticism and Calling Out in Twitter

HAESOO KIM, School of Computing, KAIST, Republic of Korea
HAEEUN KIM, College of Humanities, EPFL, Switzerland
JUHO KIM*, School of Computing, KAIST, Republic of Korea
JEONG-WOO JANG*, School of Digital Humanities and Computational Social Sciences, KAIST, Republic of Korea

Calling out, a phenomenon where people publicly broadcast their critiques of someone to a larger audience using, has become increasingly common on social media. However, there has been concerns that it could develop into harassment, deteriorating the quality of public discourse by over-punishing individuals for minor transgressions. To investigate this phenomenon, we interviewed 32 Twitter users who had been called out, had called out, or had witnessed a calling out on Twitter. We found that a key determining factor that distinguishes criticism from harassment was the callee's ability to respond to or engage with the criticism, and that different stakeholders hold different perspectives toward how online harassment is defined. We also discovered that the distinction between callers and callees was not absolute, and that there was high interchangeability of roles both within and across events. Through these findings, we discuss design implications for the platform in promoting healthy discourse while preventing toxic behavior on social media.
Content warning: this paper mentions sensitive topics (e.g., self harm, suicide) related to online harassment.

CCS Concepts: • **Information systems** → Social networking sites; Social networks; • **Human-centered computing** → **Empirical studies in HCI**.

Additional Key Words and Phrases: Online harassment, cyberbullying, social media, computer-mediated communication, social justice, Twitter

ACM Reference Format:
Haesoo Kim, Haeeun Kim, Juho Kim, and Jeong-woo Jang. 2022. When Does it Become Harassment?: An Investigation of Online Criticism and Calling Out in Twitter. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 474 (November 2022), 32 pages. https://doi.org/10.1145/3555575

1 INTRODUCTION

Since its conception, Twitter has been a space for users to share their opinions and thoughts on various social issues. People are able to communicate their interests on the platform, discuss controversial issues [117], and even participate in political discussions [7]. As these topic networks are formed, online social networks operate as a public sphere [7, 116] where various social issues are discussed through open communication [118]. While some have pointed out the limitations of

*Jeong-woo Jang and Juho Kim are co-corresponding authors.

Authors' addresses: Haesoo Kim, haesookim@kaist.ac.kr, School of Computing, KAIST, Daejeon, Republic of Korea; Haeeun Kim, haeeun.kim@epfl.ch, College of Humanities, EPFL, Lausanne, Switzerland; Juho Kim, juhokim@kaist.ac.kr, School of Computing, KAIST, Daejeon, Republic of Korea; Jeong-woo Jang, jwjang29@kaist.ac.kr, School of Digital Humanities and Computational Social Sciences, KAIST, Daejeon, Republic of Korea.

2573-0142/2022/11-ART474 $15.00
https://doi.org/10.1145/3555575

such online public spheres in facilitating true democratic conversation [26, 69], there have been cases where such conversations extended past the online space and brought significant changes in the 'real world' as well [30, 40].

In some cases, such conversations happen through critical communication. A particular method of criticism that has recently gained prominence is 'calling out': the public identification and criticism of individuals online [8, 9, 94]. Sometimes referred to as 'cancel culture' [51, 110], this refers to the public criticism and withdrawal of support for those who are assessed to have said or done something problematic, often from a social justice perspective [85]. In this paper, we use the term 'calling out' to refer to the general act of publicly criticizing someone online for a perceived transgression.

Calling out has been used for a variety of reasons, ranging from private conflict resolutions to a worldwide discussion on sexual harassment [79, 84]. However, there has been skepticism on whether this form of opinion sharing truly facilitates public discourse. Previous work has suggested that public conversation often focuses on the individual in favor of discussing high-level concepts or structural issues that may have influenced the individual's behavior [15, 95]. In such cases, morally-motivated critiques toward the individual, while well-intended, could easily progress into online harassment [49, 74].

Previous research has explored the communicative values of calling out, but there has been a lack of consideration about the role and influence of calling out in public communication. To our knowledge, there have been relatively few attempts at identifying the factors that lead users to think that they are being harassed, and not just criticized. In this paper, we aim to expand upon this subject, focusing on the various experiences surrounding a calling out and how it is interconnected with online harassment.

We interviewed 32 Twitter users who have experience with either being called out (Callee), have participated in calling out someone (Caller), or have witnessed it happen (Bystander). We discovered that Twitter users consider calling out and harassment as highly interconnected concepts, and that calling out has a high probability of progressing into harassment, especially when certain conditions are met. While critical conversation was considered an important part of social media communication, calling out was generally perceived as an ineffective approach for persuading or initiating conversation. Instead, participants noted that callers mostly used calling outs to express their own opinion, using the callee's tweet as a tool for amplification and not for conversation.

We also discovered that perceptions of what constitutes harassment differed between stakeholder groups. Callers thought that the actions of individuals involved in networked harassment [65] should be evaluated independently, while callees perceived them to be indistinguishable from the actions of the group. Through this, we provide implications on how online harassment should be defined, and how platforms might build mitigation strategies according to these competing definitions. We also identified common factors that were involved in progressing a calling out into online harassment, as well as general patterns of calling outs. Through these findings, we discovered that contextual background and prior perceptions about the subject matter play a large role in the decision to call someone out. Finally, we discuss the role of the platform in facilitating civil conversation, and suggest design implications for preventing or mitigating the effects of online harassment.

The contributions of this research are as follows.

- A descriptive model of online calling out behavior, identifying common patterns based on diverse user experiences
- Insights into the perception of Twitter users on calling out and what distinguishes online harassment from 'valid' criticism

- Design implications for preventing online harassment while encouraging healthy discourse on social media

2 BACKGROUND AND RELATED WORK

In this section, we first observe previous work on calling out behaviors on Twitter, with focus on how it is perceived by the Twitter user base. We then discuss previous research on social media justice and online harassment, and establish a clear conceptual background from which we will investigate calling out behaviors.

2.1 The Prevalence of Calling out on Twitter

We first begin with an observation of calling out on Twitter. Public criticism behaviors have been referred to in various ways, including 'cancelling' [29, 51], public shaming [95] and calling out [84]. The more common term 'cancel culture' was coined in Black Twitter, where the hashtag *#cancelled* was used to critique and share experiences related to systematic racial inequality [20]. However, these terms were often used with negative connotations, implying that it has become a trivial habit of the public [51], or even a case of mob mentality where users would simply 'attack' people [20].

As many calling out cases happen on Twitter [15, 29, 86], there have also been concerns about the limitations of the platform itself in facilitating further conversation based on the criticism. Twitter has been criticized in that it merely encourages moral outrage rather than rational discussion [16, 88]. As calling outs became prevalent, casual terms such as "Twitter's villain of the day" [68] have also emerged, implying the commonality of calling outs. These limitations have been attributed to the relative lack of effort involved in tweeting [15], the high speed with which text is disseminated [61], as well as the lack of nuance in the limited space [85]. Bouvier observed that tweets using a 'cancel culture' hashtag would often represent racism as a personal, homogeneous trait [15], instead of a systematic and complex issue that goes beyond the individual. Such tendencies have been noted to potentially distract from the social context that enabled such behaviors, reducing them to an action of the individual than a societal, structural issue [16, 42].

Despite its pervasiveness in online discourse, there has not been much research on how being called out might impact the individual. In his book, *So You've been Publicly Shamed*, Jon Ronson presented accounts of subjects of high-profile online calling outs, and of the impact it had in their lives [95]. However, there has been little previous effort to understand the motivations for calling someone out, as well as its bigger impact on the larger Twitter community. Moreover, by mostly focusing on public figures, many overlook the fact that the call-out culture has become prevalent online, subjecting ordinary individuals to high levels of public scrutiny [15].

2.2 Performing Justice in Social Media

Much previous work has highlighted social media for its potential for facilitating democratic communication, as well as bringing communities together to mobilize for social justice. Bonilla and Rosa noted that digital activism garners interest from populations that are more likely to be misrepresented by media [14]. Similarly, Salter notes that victims of sexual violence have been able to claim a more prominent position by garnering a more sympathetic public as well as authority through online channels [97]. This emphasizes the role of the internet to operate as a counter-public space [6, 97], challenging existing communicative hegemonies through consciousness-raising [37] and redemocratizing public conversation [81, 97, 98].

Calling out behaviors have been used as an attempt at restoring justice where criminal justice laws could not perpetrate [91]. As youth are less likely to trust social media companies or legal authorities to achieve fair resolutions in social media disputes [100], they instead turn to more

personal modes of intervention such as criticism [12] or a public demand for an apology [99]. Here, the act of calling out instigates social change by encouraging people to re-evaluate their previous actions, as well as creating lasting conversation on the reality of social justice [40, 79].

On a society-wide scale, social media has been considered a valid platform for performing identity, solidarity, and activism, especially for minority groups [66, 78, 105]. However, there has also been criticism on the subject of social media activism, mainly on its limited ability to promote active involvement, as well as possibly even decreasing motivation [64]. In particular, micro-political activities [112] have been referred to as 'slacktivism' [64] or 'clicktivism' [44], in that it requires low personal risk or effort while mostly only providing satisfaction to the person engaged. Others have argued that despite the low level of involvement, micro-political actions have potential to promote social engagement as well as bring substantive change to society [44, 45].

Finally, users may attempt to take matters into their own hands. While 'cancel culture' focuses on high-profile individuals such as politicians and celebrities [86], everyday individuals are also subject to such scrutiny when they are perceived to have done something wrong [56, 85]. This can be observed in a retributive justice standpoint, which suggests that individuals receive a proportional, 'deserved' punishment for their actions [114]. Blackwell et al. explored how retributive approaches are perceived by social media users in response to a perceived transgression [12]. Marwick introduced the concept of Morally Motivated Networked Harassment (MMNH) where people utilize networked harassment to reinforce social and moral norms [74]. Here, people use calling out behaviors as a form of social shaming, upholding social norms by publicly humiliating the callees [58]. Klang describes this phenomena as cybervigilantism, pointing out that callers often face no physical or emotional challenges in the process, which brings their moral legitimacy into question. [56] We aim to extend upon such work by exploring how people act around morally motivated conflicts on Twitter, and how it is perceived by other users.

2.3 Online Harassment

Networked online spaces are fundamentally different from offline, unmediated spaces in terms of its persistence, searchability, replicability, and the invisible audiences [17]. As social dynamics are altered by such properties, the dynamics of harassment also develop unique forms and challenges in online spaces. People are more prone to harassing others in online spaces than in offline [102], and some have noted that it may cause more psychological damage than offline bullying [41]. Anonymity also has a significant impact on online harassment, as it can foster disinhibition and deindividuation within users [70], reducing their sense of responsibility [27, 106] and magnifying deviant behaviors [31].

Traditional definitions of bullying include elements such as repetition of messages, power differential between the perpetrator and victim, intent to harm, and aggression [63]. However, due to the aforementioned differences in social dynamics, they cannot be applied directly to online spaces. For example, the element of repetition is extremely facilitated in online contexts as online content is highly persistent [23, 77] as well as distributed to a larger potential audience [10]. This makes it difficult to control who gets to access and reproduce harassing content. Similarly, while power differentials are traditionally based on individual power relations in offline societies, online power differentials can be caused by other elements such as anonymity and volume [63, 71, 80].

Another challenge in defining online harassment is that it is hard to reach an agreement on what actually constitutes harassment. Many users who are accused of being harassers may complain that a simple disagreement was portrayed as harassment by other users [49]. Even when the intent of a message is not necessarily to harass, it could be perceived as harassment when many users join in (referred to as 'dogpiling' [12, 47, 49]). There are also cases where online harassment is seen as justified. Blackwell et al. observed that users perceive online harassment as more justified or

deserved when the target has committed some offense [13]. Others have voiced concerns about the desensitization due to the prevalent harassing behaviors in online spaces [111].

Most social media platforms adopt some form of content moderation to protect users against abusive behavior [39], but platforms usually do not have a clear definition of what constitutes abuse [90], nor are they well-communicated to their users [62, 83]. Moreover, as online content moderation usually focuses on the criminal justice standpoint of punishing the offender [90], less attention has been made to address the impact on the targeted user [76]. Schoenebeck et al. emphasizes the importance of defining an act as harassment to provide a way for individuals to find closure or validate their experiences [99]. We extend upon this line of research by collecting the stakeholders' perspectives and present a novel factor in defining online harassment.

2.4 Research Context: Communication Features on Twitter

In Twitter, there are many forms of reacting to a tweet or communicating with a particular user. The officially supported forms of reacting to a tweet are as follows: 'likes', representative of a person's agreement or preference to the content of the tweet [113]; retweets (RT), where users directly repost messages posted by others [18]; quote-tweets (QT), where users are able to directly repost others' tweets while adding their own comment as a new tweet [38]; and finally replies, a commenting format that adds and displays the reply in thread format from the original tweet [93].

While not supported officially by the Twitter interface, Twitter users also use a method commonly referred to as Latest RT (LRT), which involves retweeting a tweet and immediately making a separate tweet in reference to 'the tweet I retweeted just now' [2]. This is often used to discuss a tweet or its contents without engaging the original tweet author, as QTs can be traced from the original tweet as well as send a notification to the author. Methods of directly engaging a user include mentions, acknowledging and alerting a user by 'tagging' them in a tweet [93], and direct messages (DMs), private messages accessible to only the sender and receiver.

3 METHODS

In our IRB-approved study, we interviewed 32 Twitter users (age M = 25.72, SD = 4.20) from the following categories: *Callee* (n = 10), those who have experience being publicly called out; *Caller* (n = 15), who have publicly called out someone on Twitter; and *Bystander* (n = 7), who have witnessed a calling-out situation happening. We included the bystander group as they could have an important role in calling out or harassment by deciding to intervene (or not). Through such decisions, bystanders have the potential to significantly influence the progression of the event [35], and therefore were considered an important stakeholder.

We aimed to answer the following research questions through the interviews:

RQ1. What are common patterns and motivations of calling out on Twitter?
RQ2. How do calling outs impact the callee, and the Twitter community at large?
RQ3. How do various groups perceive or evaluate calling outs differently?
RQ4. How do calling outs escalate into online harassment?

3.1 Participants

We defined being 'called out' as instances that fit the following criteria. To say that someone has been called out refers to a situation where: 1) the criticism directly references the individual via tagging the account, quote-tweets, or screenshots; 2) it was redistributed to an unspecified public, such as the caller's followers, or the followers of people who have retweeted or reposted the original Tweet; and 3) it was posted on a public account. This condition was applied to all three groups, and was included as part of the recruitment post. We only accepted participants between the age

of 19-65 to comply with the IRB guidelines at our institution. We however note that all of the applicants were in their 20s to early 30s.

Participants were recruited through two rounds of public Tweets posted by the researchers, stating the purpose and criteria for selection as well as an open request to spread the tweet. This was so that we could utilize the amplification networks of Twitter to reach a larger potential audience. The recruitment post was RT'ed and QT'ed over 350 times, with 93,000+ total impressions. We also note that some tweets that referenced the recruitment post gained significant attention, one of them receiving nearly 1,000 retweets.

The participant demographics are organized in Table 1. IDs indicate the primary category of participant, bystanders(**B**), callees(**E**), and callers(**R**). We note that this does not constrict the experience of each participant as many participants had experiences across multiple calling out incidents and categories. The primary category was selected by the participant at time of recruitment, where we asked them to select the experience they identified the most with.

We recruited more caller participants than from other groups due to the versatility of their experience. Many caller participants reported to have experienced being called out themselves, while not as many callee or bystander participants reported to have been a caller. We attempted to balance out the overall variety of experiences through increasing the number of caller participants. In total, 19 participants identified to have called out someone (2 from Callee group, 15 from Caller group, 2 from Bystander group), and 20 participants identified to have been called out (10 from Callee group, 10 from Caller group). All 32 participants had experiences as bystanders.

3.2 Interviews

We conducted semi-structured interviews with participants through Zoom video and audio calls. Interview sessions lasted between 48 and 119 minutes, and each participant was paid 15,000 KRW (approx. 13 USD) in compensation, with the exception of two participants who refused payment. All interviews were conducted in Korean.

The interviews started with basic background questions, including demographic (age, gender, etc.) and the participant's Twitter usage patterns. Following this, each group received different questions according to their experience. The callee group was asked about the general experience of being called out, their reactions, as well as the lasting impact. The caller group questions focused more on why they called someone out, as well as how they decided to speak up. Bystanders were asked to focus on a specific incident, whether they intervened, and how the calling out progressed after that. All participants were asked if they had experience being in a different group. The genuineness of each account was verified through screenshots or links of relevant tweets that the interviewees provided. However, it was noted by the participants that relevant tweets and accounts may be deleted after the calling out, in which case the researchers utilized keyword searches of relevant tweets to verify the calling out happened.

Interview recordings were transcribed and coded through an open coding approach. Two authors individually developed a set of themes through multiple passes of the interview transcripts. We first conducted a by-group analysis where we developed a unique set of codes for each participant group (caller, callee, bystander) to observe the differences between groups and developed themes for each of them. We then conducted a second pass with all participant data, focusing on the common themes that appeared across groups and how their descriptions of similar concepts may differ. Finally, we conducted a final pass after the theme sets have been combined. Quotes have been translated from Korean to English and paraphrased for clarity.

Table 1. Participant Demographics. Calling Out experience column does not include Bystander experience as all participants had bystander experience. Anonymity was determined based on the representative account involved in the calling out case. Anonymity distinguishes if an account is fully connected to their identity (*Not Anonymous*), only discloses some personal information (e.g. age, profession, school) (*Pseudo-Anonymous*) or if they did not reveal any personal information in the account (*Anonymous*) [57]

ID	Gender	Cisgender/ Transgender	Age	# of Accounts	Anonymity of Account	Calling Out Experience	
						As Caller	As Callee
B1	M	Cisgender	23	1	Anonymous		
B2	F	Cisgender	22	4	Pseudo-Anonymous	O	
B3	F	Cisgender	27	2	Pseudo-Anonymous		
B4	M	Cisgender	33	1	Pseudo-Anonymous	O	
B5	M	Cisgender	20	1	Anonymous		
B6	F	Cisgender	33	5	Pseudo-Anonymous		
B7	F	Cisgender	21	5	Pseudo-Anonymous		
E1	F	Cisgender	21	2	Anonymous		O
E2	F	Cisgender	31	3	Anonymous		O
E3	F	Cisgender	28	1	Anonymous		O
E4	F	Cisgender	23	4	Pseudo-Anonymous		O
E5	F	Cisgender	28	3	Anonymous		O
E6	F	Cisgender	31	3	Anonymous	O	O
E7	F	Cisgender	35	2	Pseudo-Anonymous		O
E8	F	Transgender	24	5	Not Anonymous	O	O
E9	F	Cisgender	21	5	Pseudo-Anonymous	O	O
E10	F	Cisgender	26	4	Pseudo-Anonymous	O	O
R1	F	Cisgender	21	6	Pseudo-Anonymous	O	O
R2	Does not wish to answer		21	3	Anonymous	O	O
R3	M	Cisgender	28	2	Pseudo-Anonymous	O	
R4	F	Cisgender	24	6	Pseudo-Anonymous	O	O
R5	F	Cisgender	28	10+	Anonymous	O	O
R6	Non-binary	Transgender	25	3	Pseudo-Anonymous	O	O
R7	F	Cisgender	25	2	Anonymous	O	
R8	M	Cisgender	21	2	Pseudo-Anonymous	O	
R9	F	Cisgender	22	5	Pseudo-Anonymous	O	
R10	Non-binary	Transgender	21	5	Pseudo-Anonymous	O	O
R11	F	Cisgender	27	2	Anonymous	O	O
R12	F	Cisgender	27	3	Anonymous	O	O
R13	F	Cisgender	28	4	Anonymous	O	
R14	F	Cisgender	27	3	Anonymous	O	O
R15	F	Cisgender	31	5+	Pseudo-Anonymous	O	

3.3 Position Statement

We pause here to clarify the position of the authors in relation to the current work. While we recognize the potential of democratized communication in challenging established power structures, we also claim that desensitization to potentially harassing behavior, as well as subjecting individuals to high levels of public scrutiny, could be harmful. We believe that the right to free speech and expression cannot be used to justify violating people's basic rights to be protected from abuse and harassment. We also emphasize the role of social media researchers as well as social media platforms to protect their users from abuse and ensure security.

We also note that online harassment, as a widespread systemic problem in the field of online communication, disproportionately affects women, LGBTQIA+, and people of color, among other marginalized groups. Particularly in the context of Twitter, the number of followers - or supporters - can create privileges and power structures independent of their position of society. Following the concept of intersectionality, we recognize that such power imbalances are not absolute, and that multiple forms of inequalities may combine or overlap to create unique experiences.

4 RESULTS

The results of this study are focused around four major categories. First, we observe the common patterns of a calling out based on the collective experiences of our participants. Second, we move on to how and why calling outs occur by observing the motivations and patterns of callers. Third, we review the effects and impact that the calling out had on the callers and bystanders, as well as the Twitter community at large. Finally, we compare and contrast the concepts of calling out and online harassment, identifying the distinction between the two, and the factors that influence the perception toward online harassment.

4.1 Calling Out on Twitter

In this section, we describe the common phases of calling outs and discuss the factors involved in the transition between them. We also categorize distinct types of calling out behavior, which can be applied to both individual comments as well as the overall calling out incident. However, there may be multiple callers and tweets pertaining to a single calling out incident, which may consist of various different types of behavior. We note that such volatility is a central element that needs to be taken into account when analyzing calling outs.

4.1.1 Lifecycle of a Calling Out. We propose a model that represents the lifecycle and interim phases of a calling out based on the interview insights. In general, a calling out incident follows the sequence of Background - Initiation - Amplification - Resolution. Below, we go into further detail about each phase. A summary of the overall model is depicted in Figure 1.

Background. Calling outs begin as an individual (*Callee*) displays an act or comment that is seen as deserving criticism. Often, these comments are seen in connection to a larger context within the callee's own previous actions, or the community that the callee is perceived to be a part of. Other Twitter users could have been previously exposed to such contextual information, which may have caused fatigue and frustration that further motivates one to call out a person. Therefore, the calling out is often not independent, but closely connected to the context and background that callers have already developed regarding the comments similar to the callee's.

Initiation. Once the tweet gains attention from people who disagree with the callee's words and/or actions, they (*Caller*) publicly announce their disagreement, or 'call out' the callee. This may either have a single point of initiation or have multiple independent points of initiation. This is partly influenced by the callee's pre-existing networks. For example, some interviewees mentioned

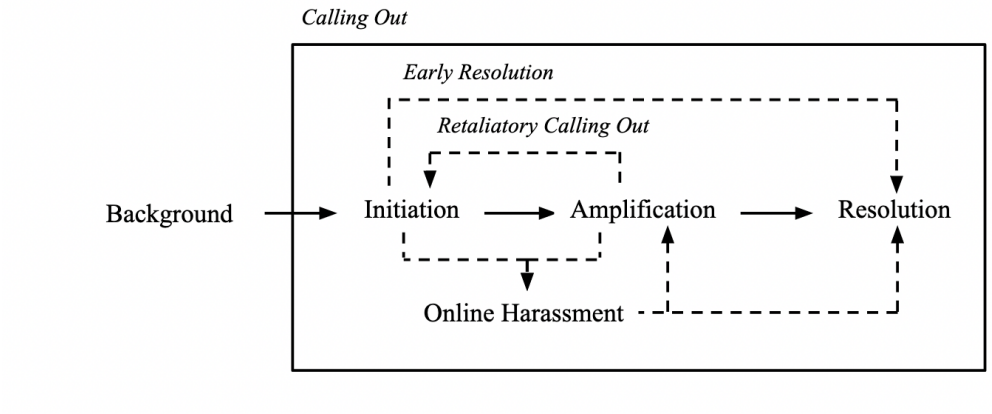

Fig. 1. Lifecycle diagram of a calling out. Solid line arrows denote the general transition between phases, and dotted arrows denote notable deviations from the central lifecycle.

that, due to their larger following, it was easy for their tweets to be noticed by others and attract criticism. Sometimes, the calling out would be initiated in private account networks, and then brought to the surface by a caller with a public account.

Amplification. In some cases, the callee may take immediate and sufficient action based on the criticism, or the callers' tweets may fail to attract the attention of a larger audience (*Early Resolution*). In others, the callee and its subsequent criticism gain further attention, attracting more potential callers and witnesses. This often rely on the interface affordances of Twitter, such as the follower-following network, topic-based recommendations, and the trending topics menu. In the process, people who disagree with the initial caller group, taking issue with the content or form of criticism, may initiate their own calling out. In this case, the callers of the initial calling out may become callees in the following calling out (*Retaliatory Calling Out*).

Harassment. During the *Initiation* and *Amplification* phases, malicious actors may begin harassing the callee through false information, vitriol, or personal attacks. Participants often distinguished between harassment and criticism according to the perceived intent of the caller, or the aggressiveness of their comments. Other cases of online harassment include situations where the scale of the calling out expands, causing psychological suffering for the callee and deterring their efforts to respond and communicate. While harassment sometimes happens unintentionally, it might also happen purposefully when the caller attempts to 'punish' the callee with the harassing responses. Sometimes, harassment may result in further amplification of the calling out as it would gain a larger audience and more people would join in to criticize either the callee or callers.

Resolution. Finally, the calling out and/or subsequent harassment dies down as callers lose motivation or interest to continue, or as callees take decisive action against the calling out. Most calling out cases are sustained through a relay network of participants. Each caller may only make a handful of comments, but each comment becomes a locus of attention that attracts further amplification. Calling outs often end when the amplification process dies down naturally and the 'flow' moves onto another subject. In other cases, callees might attempt to clarify false information pertaining to the calling out, delete their accounts/turn their accounts private, or take legal actions.

Table 2. Patterns and motivations of calling someone out

Category	Subtypes (# of cases)	Description
By Inciting Event	Organic (16)	There is no prior calling out that caused the calling out
	Retaliatory (17)	The calling out is in response to an initial calling out
By Intent	Communicative (9)	The caller wishes to engage in conversation with or expects further responses from the callee
	Non-communicative (15)	The caller does not intend to or expect to engage in conversation with the callee
	Malicious (13)	The caller explicitly wishes to harass the callee
By Intended Audience	Callee (12)	The caller is speaking directly to the callee
	Non-Callee (24)	The caller wishes to express a message to a wider potential audience

4.1.2 Common Patterns of Calling Out on Twitter. While the reasons for calling out were diverse, we observed several themes that could be used to categorize calling out events. The three major criteria were 1) Inciting Event, 2) Intent, and 3) Intended Audience. For each criterion, there were several subtypes that further defined how the calling out would proceed. The categories are organized in Table 2. We note that the subtypes are not mutually exclusive, and a caller may have had multiple motivations to calling someone out.

By inciting event, we identified two major patterns: while many calling out cases were born naturally, the behavior or actions of the callee driving the criticism (*Inciting Event: Organic*), it also had the potential to cause retaliatory calling outs, where the callee or people who sympathize with them would call out the caller of the initial calling out (*Inciting Event: Retaliatory*). In retaliatory calling outs, callers would comment on the content of the original callers' criticism ("Your arguments are wrong"), or their attitude and tone ("You cannot say that, no matter what they did"). Some participants noted that when multiple retaliatory calling outs happen in short sequence, it would no longer be perceived as a simple conflict or harassment but rather a fight between two groups or entities, opening different perspectives in its interpretation.

In terms of intent, there were three major categories. A caller could either have the inclination to converse and resolve the issue (*Intent: Communicative*), or they might not be interested in communicating with the callee at all. The latter could be further specified based on whether or not there was a clear display of malicious intent from the caller (*Intent: Malicious*), or the lack of will to communicate was simply based on disinterest (*Intent: Non-communicative*). Malicious intent was often characterized by unprompted vitriol and foul language, or threats to the callee.

Similarly, the intended audience of the caller also differed, and had an impact on how the message was constructed or delivered. In some cases, the calling out message was intended for the callee to listen directly, as a method of starting conversation or attempt at persuasion (*Intended Audience: Callee*). Other times, callers would use this as an opportunity to broadcast their perspectives or opinions to a wider audience, asserting their point of view towards the callee or the calling out (*Intended Audience: Non-Callee*).

4.2 Why do People Call Out Others?

In this section, we further discuss the motivations and actions of the callers, focusing on how and why they may decide to call out others. We also discuss how they felt about the results of the calling out, and whether they felt their initial purpose in calling out was fulfilled.

4.2.1 Motivations. Callers often discovered callees through their follower networks, where people would already be criticizing someone, as well as recommendations from their home timeline and the trending topics menu. These interface elements enabled callers to discover a calling out that was already happening, even if they were not actively searching for them. Even when there wasn't necessarily a leading calling out, high-profile tweets with many likes and RTs were also a common target of calling out due to their high visibility. In many cases, callers noted that they discovered the callee's tweets because they were already being criticized by other people, and they would end up joining in, rather than actively searching out for someone to criticize.

> Someone has to be criticizing it already for it to reach me, because I don't go looking for those opinions. - R6

One major reason for calling out was to correct a factually incorrect or misleading statement. Callers mentioned that they wanted to prevent misconceptions and potential harms that may occur due to the spreading of false information. For example, R12 called out a Twitter user for spreading wrong information about veterinary treatments.

> They were taking issue with the actions of a medical professional, and nonprofessionals shouldn't really say these things about professional treatments when they don't know better. I know because I'm in the field myself. - R12

Another motive was to signal the inappropriateness of the callee's comment. In this case, callers would use the callee's tweet as a counterexample to promote their opinions about a subject. These were mostly based on social justice topics such as hate speech toward minority groups; misogynistic, homophobic or transphobic comments; or offensive comments directed to groups such as people of a specific profession, ideological groups or even fandom. Many callers mentioned that they valued the ability to reach a larger audience through the callee's tweet. Therefore, their motivation was not to communicate with the callees, but rather to let bystanders know of the error, preventing potential harms that may occur due to the spreading of false information.

R6 mentioned that they spoke up mostly to fight against misconceptions or hate speech about their cohort, which included being a nurse and a non-binary individual. Because of this, they thought it was their responsibility to speak up to defend such minority groups.

> We're outnumbered. When I speak up, it's always from the minority's side. For us, it always helps to have someone speak up. - R6

Callers would also use calling outs, and the resulting networked reaction, to pressure the callees and people with similar perspectives to them. Calling out someone and sometimes harassing them was their way of letting others know that there will be consequences to similar actions. This also had the intent of pressuring bystanders with the implications of potential consequences, using the callee as a scapegoat. Callers noted that there was power in numbers, and they sometimes leveraged their following or follower networks to attract more people that agreed with them. In these cases, the callee's tweet was used as a vessel to convey a bigger idea to the larger Twitter sphere.

> I wanted to show my views to others by criticizing them. It has a much larger influence if I'm criticizing someone than say, if I'm writing it in my blog. So I wanted to express these views. - R14

> We have a community of nurses who are all mutuals with each other. So when I criticized [the callee] for insulting me and my job, those friends rushed to them and started demanding that they apologize. - R6

Finally, callers tended to speak up if they felt they had a unique point to contribute, such as an example from personal experience, a novel point of view, or factual evidence that had not been previously mentioned. For example, if the existing critique contained a specific type of relevant experience, callers might not choose to join the calling out since they felt their comment might not add anything unique to the discourse. On the other hand, if their initial assessment of the calling out was lacking a specific anecdote they felt would be relevant, they would be more likely to join in.

4.2.2 Leveraging the Twitter Interface. One of the most common forms of calling out was through QTs. Some participants reported to occasionally use replies or LRTs in place of QTs, but the overall consensus was overwhelmingly skewed towards QTs, and many participants mentioned that QTs were a common method for calling out on Twitter. Callers remarked that they would often use QTs instead of replies because it was often not their intention to communicate individually with the callee, and choosing such a direct mode of interaction caused additional social pressure for them. QTs were considered a more indirect way of criticism, with focus on expressing their own opinions and communicating with their own followers.

> QTs do feel different. If you're replying to them, it's like shouting to them, "Hey you!" when on the other hand QTs are like "Hey, check out what a stupid thing this person said." It feels a lot less burdensome. - R10

In relation to this, callers noted that Twitter users often use QTs as a reference to form their own opinion about a subject. Knowing this, they would purposely QT tweets that have garnered a lot of attention (both critical and favorable) and would try to take advantage of the popularity of the original tweet. For some participants, this also influenced how they would choose a specific tweet to criticize.

> It's more that I want to show this tweet, and what I think about it, to my followers. In that sense, I suppose the callee is more of a scapegoat for me to express what I want to say about this topic. It's a way to increase exposure about such subjects. - R3

> I want as many people to see my tweet, so I purposely choose the one where there's a lot of RTs and QTs to express my opinion. - R15

In some cases, the number of QTs was used as a proxy to determine the appropriateness of the callee's original comment. This is related to the idea of being 'ratioed' [3], referring to situations where there are more replies or QTs (comments) - representative of disagreeing comments - than likes or retweets - representative of agreement. Similarly, B2 noted that the perception toward QTs are mostly that they are critical, especially en masse.

> People say that if there are more QTs than RTs, then whatever you said is problematic. - B2

Some callers would go as far to use dedicated burner accounts, separated from their main account, to call out someone. This was sometimes used to avoid the possibility of retaliatory calling outs. R7 had a dedicated public account with "no profile picture, followers or following, no connection to any identity" so that they could freely talk about social issues or call out others without the potential of being called out in retaliation.

> It's an account with nothing in it, so the negative reactions to it don't really exist even if people try to attack me. - R7

R5 and R11 also mentioned that they took care to make sure that the accounts they used for calling out cannot be traced back to themselves for fear of being identified (R5) or the possibility of legal retribution (R11). They also mentioned that their followers or Twitter friends could feel fatigued from the aggressive tweets they made, which led them to run a dedicated account.

> As I grew deeper relationships with my Twitter friends, I wanted to only show better versions of myself to them. So I started to call out people on another account. - R5

4.2.3 Was the Goal Achieved? As mentioned in previous sections, most callers identified their motivations to be of some combination of persuading or correcting the caller (*Intent: Communicative*) and attempting to reach a larger audience and raise awareness about the issue or opinion by using the callee's tweet as a medium (*Intent: Non-communicative*). Callers noted that it is much rarer to succeed in persuading callees, and that callees would more often simply ignore the calling out or delete their account, opting for evasive responses.

When the intended audience was not the callee, callers would more often perceive their calling out as a success, as such calling outs revolved around the desire to express their opinion about a specific issue. However, when the motivation for calling out was primarily communicative, many mentioned that it was often unsuccessful. All callers agreed that calling out rarely ended in a successful conversation with the callee. Neither did anyone report to have had success in influencing the callee's opinion. Participant R3 mentioned that their motivations would vary for each calling out, but the communicative motivation had the lowest rate of success.

> I mean, in terms of bringing this issue to light and making it more visible, I think it works. In the persuasion front, not so much. It's much rarer that that happens. - R3

4.3 What Happens After Someone is Called Out?

In this section, we focus on the callees' and bystanders' accounts, centered around their reactions and countermeasures, as well as the lasting effects it may have had on people who have experienced or witnessed calling outs.

4.3.1 Reactions to Being Called Out. In response to a calling out, many callees' immediate emotional response was fear and anxiety. Even if the calling out was relatively small or less intense, the immediate fear of being criticized, as well as the panic that they may have potentially said something controversial was observed across many callees before they were able to make sense of the situation. As the calling out amplified and grew into harassment, callees often reported to have felt scared, and being paralyzed to the level of being unable to take action. This was especially the case in larger calling outs, where callees would be taken aback by the response. In such cases, callees reported to have been at a loss, feeling helpless from being unable to respond to the criticism. They noted that as calling outs happened, they were exposed to audiences that are much larger or different from what they had anticipated. This caused them to be taken off guard and unprepared for what followed.

> I was just posting what I thought, but all of a sudden I was the center of attention. And all of these people were being really critical. That scared me. - E3

> It wasn't that critical in the beginning. My friends all found it funny, RTing to laugh along, but then suddenly the RTs exploded and everything just escalated really quickly. - E5

4.3.2 Responding to Criticism. Response patterns from callees ranged from no response at all to legal action, and in some cases escalated as far as callees threatening to commit or actually committing self-harm or suicide. While the specific form and consequences differed depending on

Table 3. Common response patterns of callees

Category	Response Type	Description
Passive Response	No Response	Callee does not acknowledge that they are being called out, or interact with callers.
	Deleting Tweet	Callee deletes the tweet that is being criticized or called out.
	Turning Private	Callee turns their account private to prevent other Twitter users from interacting with them.
	Deleting Account	Callee deletes their account or creates a completely new account.
Active Response	Refutation	Callee refutes the points made by the callers, either directly engaging with the callers' tweet or indirectly.
	Public Amendment	Callee posts a public tweet containing an apology or amendment of what they said previously.
	Legal action	Callee sues, or implies that they will sue, the caller(s).

the situation, there were several common approaches that callees would take. This is organized in Table 3. As callees' perceptions of calling outs were mostly aggressive, their responses also often took a defensive stance.

Many participants noted that active responses (e.g. public apologies or direct refutation) could make things worse. Callers would often take issue with the peripheral elements of the callee's message, such as tone or attitude. In particular, many participants noted that apologies would often be ignored, gaining less attention than the initial tweet or calling-out tweets. Participant E2 shared their experience regarding futile apologies.

> I did post an apology regarding what I did wrong. But people wouldn't listen, and I just got criticized more because I didn't delete the original Tweet. [Another person] left Twitter after apologizing and deleting their tweets, but people would still keep talking, saying that it's irresponsible to just run away. My hands felt tied - What is it that they want? - E2

However, this did not always mean that passive responses were a better approach. Participant E1 experienced this firsthand when they initially tried to ignore the calling out, but it ended up backfiring on them.

> At first I thought that no response would be the best approach, so I let it be. But then I woke up to literally hundreds of notifications. - E1

Participants E1 and E2 had attempted to report the harassing Tweets, but found it unsuccessful. They noted being frustrated by the lack of response, as well as the time delay before actual interventions would happen. This caused our participants to think that the act of reporting itself is meaningless.

> I tried to stop it before more people saw it. I think I reported the account like 10 times... but nothing happened. [Twitter said] it doesn't go against community guidelines, but I feel if they paid attention the first time I reported it, this wouldn't have happened to me. - E1

Some participants also noted the dangers of the report feature being abused as a harassment tactic.

> I used to think that the report feature could be a solution to this, but then I realized that could also be used for harassment. Like a group of people intentionally reporting everything someone says so that they will be suspended. - B6

In most cases, these attempts were unsuccessful in resolving the calling out. Rather than response tactics, the scale of calling out and the escalation level were deemed more critical in determining the effectiveness of a response. If it was resolved before it could escalate, active intervention was perceived to be appropriate. Otherwise, many pointed out that it is unsuccessful or even counterproductive, as it would only cause the conflict to further escalate.

Finally, some callees mentioned that they purposefully did not take evasive action as they did not want to feel like they were 'losing'. In this case, they perceived the calling out as attacks, or even as a competition between themselves and the callers. In this case, they mentioned that using evasive tactics such as blocking them or turning private felt like giving in or admitting defeat to the callers. This attitude of resistance would also often lead to retaliatory calling outs, which could potentially reverse, or level out, the power relationships between the caller and callee.

4.3.3 Lasting Effects on User Behavior. Many callees reported to be discouraged from calling out after their experience. They empathized with the psychological pressure that callees feel when they were being cornered by multiple people, and they did not want to have another person go through a similar state. Several participants reported that this also affected their everyday lives. Participant E2 shared their experience of feeling isolated.

> My real-life friends don't know about this incident or my Twitter account, I had nobody to talk to. - E2

Their perceived efficacy of calling out someone also took a turn to the negative, as they had experienced the futility of trying to convince someone via a Twitter conversation. Participant E4 also noted that even if the criticism is valid, it is likely to be redundant, which reduced their willingness to call out someone.

> I quote tweet a lot less because I figured my input is not going to give any novel insight, but only fatigue towards [the callee]. - E4

Out participants also reported that being called out or witnessing a calling out often discouraged them from using Twitter, or at least influenced how they used Twitter. For callees who had experienced being a caller, being called out discouraged them from calling out others as they empathized with the anxiety or pain the callees might feel. Some users even deleted their accounts or moved their account to start from scratch as callers would persistently follow them and continuously re-ignite the subject. Some even reported to have left Twitter temporarily following the calling out due to the emotional toll.

In particular, many callees reported that this affected how they leverage their private and public accounts. Most of our participants had reported to use both public and private accounts: private accounts were used mostly for talking about private subjects or opinionated issues; topics they did not feel comfortable posting in their public accounts. Calling out had an effect on the use of private accounts as many participants noted that the reason why they used separate accounts in the first place was the potential of being called out. They feared the possibility that their personal information would be used as fuel for harassment, discouraging them from using public accounts as much. Similarly, callees reported to monitor what they say in their public accounts much more closely after this experience, talking less about 'controversial issues' that may attract callers.

> I just stopped saying anything that people might disagree with. I used to be really vocal about a lot of things. Feminism, politics... I just kind of moved away from talking about those things. Even seeing them became too stressful, so I often just mute[1] those topics. - R4

4.4 How does the Twitter Community Assess Calling Outs?

In this section, we discuss user perceptions and assessments toward calling out, and what factors were involved in it.

4.4.1 Online Karma: Private Realization of Justice. The perceived validity of the criticism, as well as the initial transgression from the callee, was a critical factor in assessing calling outs. Even though callers were often aware of the emotional toll they might put on the callees, they would still feel that the calling out was necessary. Their comments were mostly made 'despite the fact' that such negative repercussions exist, especially when their motivation was to prevent a potential larger harm that may come from the callee's statement.

> A lot of the comments were pretty mean, or ridiculing [the callee]. But I still think that was deserved. And other people were too, RTing or QTing some of the funnier tweets that were making fun of them. - R13

R12, in particular, described it as 'Online Karma': implying that the callees were getting what was deserved. With such different contexts and levels of transgression, the validity of the calling out was perceived differently.

> I don't think this is online harassment; it's more like online karma. - R12

4.4.2 A Tool for Public Discourse. Generally, there was a widespread agreement among our participants that calling out is still a form of public discourse and opinion sharing. Some participants also noted that calling outs, and the subsequent conversation that may be prompted from it, can still be meaningful.

> Even if it starts maliciously, [the calling out] did open the door for a lot of active discussion about the topic. - R7

For this reason, several participants were skeptical about the idea of moderating or regulating calling outs, even as they acknowledged the possibility of harassment. They expressed concern that the open communication model of Twitter could be compromised if too many preventative measures were taken. However, even as participants recognized the value of calling out in there were differing opinions about how appropriate or effective it is.

4.4.3 Limited Tangible Effect. Many participants, especially those with callee experiences, expressed skepticism on the communicative value of calling outs. They noted that these calling outs rarely had the effect or intention of persuading the callee. Participants also noted that calling out is an increasingly common phenomenon in Twitter. One repeated sentiment was that it "happened too often to remember", implying that the prevalence of calling out behaviors was such that individual events became indistinguishable from one another. Many participants reported that they witnessed similar events multiple times on a weekly or even daily basis. This caused them to be desensitized, and callees would often choose to not acknowledge criticism even when they were called out.

Interestingly, a similar atmosphere of skepticism was observed even in the callers, despite their involvement. However, their feeling of skepticism was more connected with the fatigue coming from their calling outs failing to persuade the callee. Therefore, they would be discouraged from attempting to reason with or communicate with the callee, and simply move immediately

[1] Muting a keyword prevents all tweets that contain that keyword from appearing in a user's Home Timeline

into non-communicative calling outs. Unsatisfactory results would lead to callers experiencing fatigue regarding the efficacy of their involvement. In some cases, it even resulted in shifting their motivations behind calling out or calling out people less in general. Many callers mentioned that their motivations for calling out turned from communicative to non-communicative as they realized that their efforts at reaching out to the callees were often ignored.

> There are so many people saying things [that I don't agree with], but there's no end if I attack each one of them, and they'll all just come back to attack me again. It's an endless cycle. - R11

4.5 How do Calling Outs Escalate to Harassment?

In this section, we discuss how perceptions of online harassment might differ between stakeholder groups, as well as what the heuristic standards that Twitter users utilized were to distinguish cases of harassment from calling out cases.

4.5.1 Factors that Constitute Harassment. Here, we point out the various factors that participants mentioned that transitioned a calling out case into online harassment. While some factors (*Spreading Wrong Information*, *Determined Following*, *Vocabulary and Tone*) focus on the intent of the caller, we note that other factors (e.g. *Scale of Comments*) were independent of the callers' intent, which opens up the possibility of 'unintentional' harassment.

Scale of Comments. Many participants agreed that the sheer volume of comments would often could cause a feeling of fear and being overwhelmed for the callees. Even for bystanders and callers, the scale of the calling out had an impact on how they would perceive the calling out. When it was larger, they would often feel more sympathetic towards the callee and perceive the event as harassment. This in turn had the effect of transferring comments with communicative intent to non-communicative in nature, as the callee was not able to process or engage with it all. Some callers reported to have been discouraged from calling out a tweet they perceived as wrong because there was already many people criticizing the callee, worried that they might cause harassment.

> I think it's also harassment when it becomes so big that everyone starts chipping in. I look at the gravity of the situation, and if there's like, thousands of QTs already I just pass by without saying anything. - R7

Callees also reported that if a calling out grew in scale, it would cause them to be overwhelmed by the situation, which influenced their ability to react or respond to the criticism of the calling out. Many callees noted that as the scale of the calling out grew, they were unable to read or interact with all of the messages, and their ability to communicate was lost to them. This meant they were also deterred from trying to clarify misunderstandings or false information, unable to regain their autonomy in the discourse.

> I woke up and there's a notification on my phone saying that I have 99+ notifications. There's more mentions and QTs in my notifications than I can dream of. I immediately thought, ah, I'm being harassed here. I was so terrified that I couldn't even open my DMs. There were so many, I couldn't even see how many there were, let alone read them all. - E2

Participant E6 noted that in such occasions, positive or supportive comments also lose their value as it is hard to distinguish them from the malicious or criticizing comments.

> You don't know how much of it is malicious, and how much of it is actually agreeing with you. You can't read them all anyway, so it doesn't matter. But not knowing the ratio gives you even more fear. - E6

Participant E1 also shared that positive comments, while posted with good intentions, could also function as a locus of engagement, ultimately bringing more negative attention. It was also noted that people who post supportive comments for callees would also be harassed, as was the cause for being called out for some of our participants. For them, "all attention was bad attention".

> In a way, I started resenting everyone who'd give attention to the issue. Even in cases where they'd try to defend me, the people attacking me would also find an excuse to go and attack them as well. And because there's a difference in scale, there's no way we win. So all it does is make the issue even bigger. All attention was bad attention. - E1

Spreading False Information. A common element mentioned to shift a calling out to harassment was when callers would start spreading false information about the callee or begin purposefully misinterpreting their words or actions. This included exaggerating the callee's words or intent to vilify them, or taking them out of context. R5 had experience being called out for sharing a roleplay scenario set in the World War II era that involved fighting against Nazis.

> People went insane after just reading that it features Nazis. And if they listen just for a second, they'd know it's not what they think. But then everyone would just go instinctively like, 'you're pro-Nazi then'. - R5

It was noted that such comments would be fabricated to make the callee seem like a 'bad person', providing moral justification for their harassment. Other cases included incidents where the callers' personal interpretation of the callee's words or actions would circulate as if it were the actual thoughts of the callee. Several callers also agreed that such behavior is inherently malicious and undermines the validity of the calling out.

Aggressive Vocabulary and Tone. Another major factor was the vocabulary and tone used by the callers. In particular, there was an emphasis of the use of profanity or insults when calling someone out, which many users interpreted as 'counterproductive' and as 'refusal to communicate'. Even when there were no direct insults, the tone of conversation was deemed important in deciding what is harassment or not, being indicative of the perceived willingness of the caller to engage in conversation.

> You can be critical, but when it moves on to mockery or downright attacks then you know what they want is to harass you. - B3

Some participants did note that this is subjective, and identified challenges in accurately interpreting the intent of the caller. Since tweets are short and ephemeral by nature, it becomes harder to integrate nuances and context in them. This has the potential to cause cases of misinterpretation and also opens up the possibility of callers avoiding responsibility, claiming that it was not their intention.

Determined Following of Callers. Another tendency was that when the calling out was perceived as harassment, the callers would often focus more on the individual behind the account rather than the inciting actions. In some cases, this took the form of callers determinedly searching for past tweets or personal information of the callee to find more things to criticize. Many participants noted that when the callers would start to comment on unrelated information such as the callee's personality or what tweets they interact with, rather than the criticized actions itself, it would feel more like harassment than criticism. Participants noted that they felt the callers were only looking for excuses to validate their harassment and wanting more people to join in on the criticism/harassment. In the case that B6 witnessed, this went as far as hacking into the callee's private account.

> [The caller] hacked into [callee]'s private account and turned it public. So they wanted it to seem like [callee] was a bad person. And then other callers would go to [callee] and demand explanations about those things posted on the private account. - B6

Similarly, participants noted that some Twitter users would harass users by using private accounts and pressuring them into the perception of being criticized. As tweets made from private accounts cannot be seen by other Twitter users, the contents of these tweets are not accessible to callees. However, as Twitter still aggregates them as part of the total reactions to the tweet, callees and other Twitter users are able to know that the QTs exist. This caused anxiety for the callees, and Twitter users in general, as they were given the impression that they may be criticized or denounced in those QTs but they had no way to disprove this idea. This fear was partly confirmed by the accounts of the callers, where many reported to have discovered the callee's tweet through previous critiques made in their friends' private accounts.

> For private QTs, I know they're there but I have no way of knowing if they saying good things about me or bad things about me. The not knowing makes me really anxious about them. - E4

Several participants categorized this as another form of harassment as it was considered as purposeful intimidation. In some cases, as in the case of R7, they sometimes purposefully evoked this effect based on the knowledge of such perceptions.

> I do it sometimes when I want to pressure them. Because [having private QTs] makes you feel anxious, right? So when I see tweets that are just plain stupid, I just QT them, no content, just literally write "quote", using my private account. I figured they'll be curious about it, and also scared that they're being criticized. It's threatening. - R7

We add that while the interview questions focused on the experience on Twitter, some participants noted that it sometimes migrated to spaces outside of Twitter. Participant E9, who had been called out for using a bathroom that matches their gender identity as a transgender individual, had their tweet posted on external spaces including school community website as part of the subsequent harassment. The callee was faced with more transphobic comments, and began to fear that they could be outed to the school community. Such as in this case, several participants mentioned that the repercussions of the calling out will follow the callee outside of Twitter, and regardless of if the account was deleted or kept. Therefore, the harassment and its implications could not only be determined by its impact on Twitter.

4.5.2 Differing Definitions of Online Harassment. While the idea that calling out could develop into harassment was more widely accepted, one important distinction was that callers and callees would have different definitions as to what constitutes harassment. Callees often perceive calling out or the subsequent harassment as a whole, and do not - or are incapable of - distinguishing between the value of each individual comment. Therefore, they perceived the entire incident as harassment when some comments progress to have harassing quality, even if they were not all malicious.

On the other hand, callers often perceived comments to be individual, and evaluated them as such. Several caller participants would evaluate their actions differently as personally not having participated in harassment even though some others with similar opinions to them might have made harassing comments. This was also noted by the callees, who sometimes mentioned that they thought that their callers would not think they are participating in harassment.

In some cases, callers felt that their participation in harassment was justified in a self defense logic if it was caused by the callee's attacks to the callers' person or identity in the first place, such as hate speech toward minority groups.

> I do feel like I'm harassing them sometimes. But even if that's harassment, they've also attacked me, my identity, and values that are important to my survival. So if they're trying to harm that, isn't it fair for me to attack them in return? Sort of like self defense? - R5

5 DISCUSSION

In this section, we discuss the implications of our findings on calling out, online harassment and on social media in general. Based on these concepts, we also suggest possible design implications to mitigate the issues that we have identified in the current study.

5.1 Implications for Discourse on Social Media

Social media allows for open discourse and communication across a variety of topics as users are exposed to experiences they may not have been able to access before. Twitter, in particular, has high potential to host previously misrepresented topics due to its open communication model and penchant towards amplification [14]. In this way, Twitter has been used to reverse the power dynamics of media through public sympathy and functioning as a counter-public space [97]. However, as such calling out behaviors become prevalent, we argue that it may harm the Twitter community as it opens the possibility to limit conversation - leading to the platform operating not as a space of conversation but as one of hostility. In this section, we discuss such effects of calling out behaviors in social media discourse, and suggest how to mediate such effects.

5.1.1 Limited Communicative Value of Calling out. Through our findings, we discovered that while the motivations for calling out were diverse, callers often focused on being able to communicate to a larger audience than direct communication with the callee (Section 4.2.1). Participants mentioned that if their intentions had been to correct or persuade the callee, they would have used more private forms of responses such as private direct messages and replies. In this, we can assume that one of the main factors that drive such a public method of resolution is the concept of imagined audiences playing witness to the event [17]. Calling out can be seen as a case where the potential to reach a larger imagined audience is perceived to be more important than the communicative value or repercussions toward the real audience (callee).

Twitter, in particular, is a space in which the concept of the imagined audience is heavily emphasized [17, 75]. However, as there are no clear cues that clarify the size of this imagined audience, there are often misconceptions about exactly what audience they could reach from their posts [10]. Based on the accounts from our callers, we can interpret their willingness to "make more people aware" of the issue as stemming from being conscious about the imagined audience [19, 74]. However, as this happened, callees were not active stakeholders in the discourse but were relegated to a vessel through whom the conversation is raised and activated. While this has its own value in terms of facilitating public conversation, it disregards the impact to the callee. Based on this, we argue that calling out behaviors should be interpreted in a lens of public discourse, and less in the perspective of individual criticism.

5.1.2 Alienation of Callees Through Amplification. As the critique is exposed to a bigger audience in the amplification stage, callees become akin to a public figure during the duration of the calling out. They are exposed to a larger body of Twitter users, most of whom they do not have prior relationships with, and they easily become objectified as an abstract 'bad' [32]. R10 compared this to the more common phenomena of celebrity bashing.

> People might be more prone to bash celebrities while they won't do that to their acquaintances. Since you can't see those people on Twitter, they become like an abstract

> public figure. You don't know what kind of person they are, and now they're just like a game character than an actual living person. - R10

This implies that, as the callee's tweet becomes its own entity, it also makes them an abstract concept that is no longer a person and just an idea that they may agree or disagree with. Thus, amplification can decontextualize the callee, alienating them from the conversation. To mitigate this, platforms could explore the idea of priming users to the person's individual contexts in addition to the message, facilitating better understanding and more empathy between users [54, 60].

5.1.3 Impact on Willingness to Speak Up. The fear of potentially being harassed and called out turned Twitter users to be more conservative of what they express on Twitter. Participants also noted that witnessing or experiencing calling outs led them to be less likely to speak up in public, and would turn to talk only in their private accounts even if they had opinions about a subject. As people tend to gravitate towards private discussions, it might further lessen the potential of public communication regarding constructive criticism or other messages.

Moreover, as further engagement was either considered futile or counterproductive, users took an evasive attitude, such as ignoring, not engaging with callers, or just 'going private'. Bystanders also became less likely to intervene due to the perceived futility of engagement. Many bystanders feared the possibility of being harassed when they intervene, discouraging them from actively standing up in the face of harassment [35]. Moreover, as several participants mentioned, sometimes bystander action does not help but only makes things worse by exacerbating the scale of the calling out or harassment [50]. Considering the importance and effectiveness of bystander intervention for mitigating online harassment [12, 21, 22, 34, 50, 96], we suggest that platforms should design for possible bystander involvement without fear of such repercussions. This is discussed in further detail in Section 5.4.3.

Finally, the commonality of calling outs in Twitter has the potential to desensitize users to harassment, such as in the case of B6. They evaluated their experience being called out as: *"this was nothing, I knew what real harassment was like. (B6)"*. This may make users unwilling to identify as victims to harassment [13, 111] and being resigned to the possibility of online hostility [24, 92, 101]. This could potentially deteriorate the quality of discourse in online spaces, as people would have less positive expectations about communicating with others, and would be less likely to speak up in open spaces. This can undermine the ability of true victims to speak up about the damage done to them, losing chances for reparation and support. Thus, we emphasize the importance of providing safe spaces where users can freely disclose and define their experience of harassment without fear of being judged [13].

5.2 Platform Dynamics in Calling Out and Harassment

While we have mostly focused on the motivations and actions of the user in calling out, we did find that many of them were mediated by the platform affordances of Twitter. This ranged from the users' perspective of what each feature would imply in communication, as well as features that were seen as directly encouraging calling out or harassment. In this section, we detail the role of the platform in shaping the calling out phenomena.

5.2.1 Forming Distinct Sub-communities. A common experience from the callees, and sometimes even callers, was that they were presented with an audience composition that had not anticipated. The discrepancy between their imagined audience and actual audience caused users to be confronted with much bigger consequences than they had predicted. This could be attributed to the tendency of Twitter, and social media in general, to encourage selective exposure through curated timelines [104]. Sometimes referred to as the 'Filter Bubble' [89], this is perceived to have a significant influence in

how each user perceives the world. We can assume that as Twitter users gravitated towards similar individuals and form networks within their community of like-minded people, they became less aware of the heterogeneous networks that might still be reached in a few steps.

These behaviors imply the effects of the polarization of communities have had within the general Twitter space. As norms and cultures differed between groups and topic clusters, so did the implicit rules of each community and what was considered correct or acceptable. This has been observed in previous research about polarized communities, where such communities may develop very different social standards and norms [11, 46]. Such competing norms would leave a narrow window for what is commonly acceptable in society (in this case, Twitter) as a whole. Future work may focus on how such rules are developed based on a large-scale network analysis of Twitter users. While there have been multiple attempts at network analysis using Twitter follow networks in different topics [43, 67, 107], as well as the discussion of how shared behavior is developed within such groups [82], there is a lack of attention towards how these behaviors differ across groups and what might happen if these clusters collide.

5.2.2 Limitations of Response Measures. In terms of countermeasures to harassment and calling out, our participants leaned towards methods that they can take individually, and relied less on the platform. In particular, participants noted that reporting or blocking harassers was often unsuccessful, especially as the calling out grew in scale, confirming the findings from previous work [13, 49, 71]. Participants emphasized that the practicality of the report feature was undermined by the fact there was a time delay between filing the report and the corresponding action, by when it was too late to stop the escalation. Borrowing from the accounts of some participants, this may imply the existence of a critical period for preventing over-escalation, which could emphasize the importance of immediate responses in content moderation.

Moreover, as existing response measures focus on deterring the individual accounts, it becomes more difficult to protect the callees against persistent efforts such as creating new or dedicated accounts for harassment. As studied in Nova et al.'s work on Facebook's visibility controls, such low levels of identity persistence counteracts and even undermines the use of the content moderation tools [87]. We note the necessity of supplementary features such as preemptively blocking new accounts from someone [5] so as to mitigate the limitations of account-based moderation measures.

5.2.3 Amplification Features Promoting Harassment. Some participants noted that the Twitter interface had the potential to exacerbate or cause harassment through amplification interfaces. For example, when there are multiple callers, provocative or more violent posts could gain more visibility based on its engagement levels [53], dominating the conversation and enabling further harassment. As users are more exposed to aggressive reactions, they be desensitized toward them and consider such actions as acceptable.

Many of our participants noted the hostile perception towards QTs, as well as their tendency to be used with more aggressive or uncommunicative intent. While Garimella et al. have previously noted that quote tweets were less likely to be aggressive compared to replies [38], the results of the current study imply that QTs could be perceived to be more aggressive when used for calling out purposes. We note that this change in perception may be influenced by the introduction of the QT timeline [109]. The QT timeline interface was newly deployed in September 2020, allowing users to access 'tweets about a tweet' at once. With the introduction of this feature, it becomes easier for callers to potentially cultivate an atmosphere of criticism surrounding the callee, as everyone with access to the callee's tweet can also access the QTs easily. Many participants noted to used this feature to assess the callee's original tweet, and in the case of callers, see what kinds of previous critiques have been made with regards to the callee.

5.2.4 Visible Engagement Metrics. Some forms of harassment relied on the fact that the engagement numbers were visible, while the content of the engagement was not always available. Previous research has shown that public social media engagement metrics can serve as bandwagon heuristics that influence how they feel about a certain issue [55]. In relation to our findings, we suggest that the engagement metrics such as number of likes, RTs and QTs supported by the Twitter interface may impact how calling outs progress into harassment. Similarly to the hostile perception surrounding QTs, the engagement metrics and the implications of the numbers were also influential to the perception of the content.

For example, our participants reported that as QTs were a critical factor in calling out; a larger number of QTs were generally associated with the tweet being problematic. As these heuristics develop, people may brashly judge the content of a tweet, possibly developing unfavorable preconceptions without even processing the tweet on their own. In addition, many participants noted the use of private QTs and replies as a tool to psychologically corner the victim by giving them the feeling of being criticized where they cannot see. Here, there is a clear discrepancy between the implied amount of content (displayed number of comments) and information provided (number of visible comments). This inconsistency can cause anxiety to the callees, as well as more ambiguous heuristics for bystanders. Based on this, we argue that there is a need to provide more consistent information to the user, where the amount of information that they expect to see should match the actual amount of information available.

5.3 Extending the Definition of Online Harassment

Based on our findings, we suggest that the harassment phase of a calling out can be understood as a form of retributive harassment, where harassing people who have committed some offense is considered justified [12, 33]. Based on our findings, we discuss additional challenges in defining online harassment and emphasize the impact of the callee's ability to engage in categorizing a calling out as such. We expand upon Marwick's MMNH (Morally Motivated Networked Harassment) paradigm [74] by identifying contextual elements such as the background context prior to the calling out. We also enrich the schema by identifying the various elements that influence the transition between phases, identifying the diverse outcomes that may result from a calling out, even when it does not necessarily end in harassment.

5.3.1 The Role of Context in Calling Out and Harassment. Previous work on the retributive justice perspective of online harassment has discussed the impact of prior transgressions from the harassment victims, and how that impacts the perception of if the harassment is acceptable [12, 32, 33]. However, such previous research focuses on the context of the *individual* and did not consider the more complex elements that may impact the perceived justifiability of the action. Our findings suggest that the overall attitude or prior experience surrounding 'similar people and events' was a major factor in calling out, as depicted in our suggested model of the calling out lifecycle (Section 4.1.1). Therefore, the perceived transgressions were not considered only of the individual, but including the emotional fatigue that previous similar actions had had on the callees. Based on this finding, we emphasize the importance of viewing retributive harassment in a broader context, and that such previous context could be a significant motivating factor for initiating retributive harassment.

5.3.2 Unintentional Harassment. Much previous work in the field of defining and preventing harassment uses malicious intent as a key element [63, 103]. However, our findings suggest that harassment can also happen unintentionally. This insight also aligns with the results of previous research, which showed that the perception of harassment was formed independently from the intent of the speaker [47–49]. Moreover, as callers and callees differed in their scoping of harassment

(Section 4.5.2), it becomes even more challenging for callers to prove the damages that have been done to them.

Since policies and social norms also influence people's actions and their perceptions around those actions [36], many users may also be unaware of the implications or consequences of their networked harassment behaviors: thinking that as it is not punished, it is acceptable. Moreover, the commonality of such aggressive content online may desensitize users, leading them to frame such events in terms like 'drama' rather than to label it as harassment [73]. However, it may be still unfair to punish individual commenters within the network as their individual contributions may have not been significant or ill-intended on its own [99]. In light of these findings, we propose employing experience-centric paradigms in mitigating social media online harassment. We discuss in more detail in Section 5.4.1.

5.3.3 Interchangeability of Roles. Many interviewees identified to have been in multiple positions in calling out incidents. A significant proportion of the caller participants reported to have had some form of experience being called out. Cheng et al. had previously observed that while there were innate qualities that were more likely to prompt antisocial communicative behavior online, situational variables had a significant effect as well [25]. We confirm their findings on the situational quality of aggressive online behavior, while noting that anyone can easily become a harasser or victim in the same manner.

Furthermore, the existence of retaliatory calling out incidents also demonstrates that the division between stakeholder groups is highly situational and flexible. The open communication design of Twitter allows these calling outs to be chained, sometimes even escalating or reversing the flow of events. This causes further complications in evaluating the morality of each action, as was seen in the interviews. Is it okay to harass someone if they had already attempted to harass someone else, or yourself? We recognize that these relationships can be defined dynamically, and there needs to be further discussion about how harasser-victim relationships can be formed in open online spaces.

5.3.4 Ability to Engage as a Defining Factor of Online Harassment. Our findings indicate that there are mismatches between stakeholder groups and users in terms of what defines harassment, especially between callers and callees. Callers would employ an individualized model of harassment, focusing on if they had specifically displayed harassing behavior, while callees would focus on the experience as a whole, not distinguishing between individual harassers or callers. This can be considered as an issue stemming from the difference in perception toward dyadic harassment and networked harassment. Traditional definitions of bullying refer to the concept of dyadic harassment, where the focus is on the individual that harasses another. This is defined by the relationship and power dynamics between the individual harasser and victim, as well as the intent of the harasser [65].

Harassment stemming from calling outs takes the form of networked harassment, where individuals are harassed by a group or network of people on social media, regardless of the intent of the individual within those groups [49, 65, 74]. Many existing social media platforms employ the dyadic model of harassment in content moderation, focusing on malicious individual acts such as stalking, abuse or attacks, threats, and so on [90]. However, as we have seen from the results of our study, there is little support against networked harassment despite its negative repercussions to the callees.

We argue that a critical factor that distinguishes harassment and criticism in a calling out is *whether or not the callee maintains the ability to respond and engage.* For example, if a callee is unable to engage in conversation either due to the scale of messages, or because the callers do not allow room for conversation, it could be considered as a case of harassment. We recognize that this is not the only factor that defines harassment, and that additional, undiscovered factors may still

come into play. We however note the importance of introducing such factors in defining online harassment so as to better characterize and protect users against it.

5.4 Methods for Preventing Online Harassment

Based on our findings, we expand upon the previous scholarship on methods for mitigating online harassment. We propose three possible directions for preventing online harassment: introducing an experience-centric paradigm of online harassment, designing for de-escalation, and providing indirect routes for bystander intervention.

5.4.1 Employing an Experience-Centric Paradigm of Online Harassment. One of the biggest challenges of online harassment is that it is difficult to define. Most social media platforms do not clearly define what constitutes harassment even while claiming to filter them [90], nor do users agree with these decisions [49, 111]. Marwick had previously noted the importance of moderation methods that go beyond examining individual content pieces, as the same content could be considered harassing or non-harassing depending on the context [74]. Our participants also mentioned the ambiguity of whether each individual message could be labelled as harassment if the intent was not to harass, or if the harassing effect came from the collection of messages instead of the individual comments.

In light of this, we suggest that social media platforms adopt an *experience-centric* paradigm of online harassment. Instead of focusing their efforts on punishing offenders and determining whether a content is abusive, we argue that more resources should be allocated to protecting the targeted user. This will mean that the harassment will also not be determined by the content value of each post, but by the subjective experience of the victim. We believe that introducing such paradigm could significantly mitigate the issue surrounding online harassment, and provide scalable and lasting change to improve the social media experience.

For example, if a user claims that they are being harassed, platforms may employ methods to protect and distance them from the potential harassers. This might include preventing harassers (or bystanders) from finding or interacting with the victim by reducing visibility of their posts or profile. The traditional punitive model where harassing posts or harassers are taken down by the platform requires a time delay where moderators confirm that there was indeed harassment. However, experience-centric models could focus on allowing harassed users to apply interventions by and to themselves, providing them with immediate power to control the situation. This enables platforms to react quickly as it reduces the need for verification, while also reducing the potential negative impact of false reports.

5.4.2 Designing for De-escalation. Scale is a critical factor in determining whether a calling out becomes harassment. We suggest a framework of designing for de-escalation as a method of mitigating the negative impact of online harassment. This could be done through automated detection of harassment [52, 72], where the response scale, as well as the users it reaches compared to the average, could be used to temporarily 'lock off' the post. Using such methods, it could prevent further reactions so that the responses to a single tweet do not get out of hand. Another method would be to summarize the content of past discussion to prevent repetitive arguments [59], allowing the callee and potential callers to have better ability to parse what has been said. Such methods can be used to prevent calling outs from going out of control, and to keep the discourse at a more organized level.

Improving the user's level of control over their audience is also a possible approach. Currently, Twitter's features allow users to control the audience in a batch using the "Protect tweets" feature, or individually by blocking each user. However, this is not a scalable approach, nor does it allow for protection against individuals dedicated to harass. The recent Twitter feature allowing users to

control who can reply to their tweets [108] could be useful at a larger scale, such as applying the same amount of control over RTs and QTs. Other social media platforms such as Instagram [4, 5] and YouTube [1] have experimented with giving more control to users regarding engagement metrics, which could also be utilized by limiting visibility of these metrics to other users.

5.4.3 Providing Indirect Routes for Bystander Intervention. Previous research on bystander intervention focuses on the bystander effect and diffusion of responsibility [115] as reasons behind the lack of bystander intervention. However, our participants also identified the feeling of fear of being called out, as well as the possibility of further escalating the calling out as factors. As in the case of the calling out subtype *Inciting Event: Retaliatory*, bystander intervention could begin another calling out, with potential to become harassment.

With this in mind, we propose the reinforcement of indirect bystander intervention methods, as well as better integration with the platform, to encourage active intervention towards harassment. One prominent example of an indirect intervention method is the reporting feature. For example, by increasing the transparency of the report process, platforms can provide higher report efficacy to the users and encourage bystanders to intervene [115]. Another approach is to improve the categorization used in reporting. Many platforms use predefined categories of inappropriate behavior in the reporting process, which may not match up with the user's perception [28]. This mismatch of expectations may prevent users from reporting the content even if they think a post is inappropriate. Moreover, as harassment tactics change and evolve, such approaches may not be inclusive. In light of this, platforms could allow users to specify the harassing element in the post, instead of flagging the whole post. In this way, platforms could increase the specificity of reports, allowing for fine-grained intervention methods and increased perceived efficacy for the bystanders.

In addition, we suggest that using friendsourced moderation to monitor messages or posts directed to a user could be a useful way to mitigate the direct effects of harasment [71]. Select bystanders approved by the harassed user could provide a buffer from the harassing messages, allowing for scalable filtering of malicious content. Support groups, as demonstrated in systems such as Heartmob [13], have also proven to be successful. Integrating such community support features into the platform could provide emotional and technical support for the users. Harassed users will be able to suffer less from dealing with the issue on their own.

5.5 Limitations and Future Work

While we were able to examine the differences in perception across different groups and roles, each individual experience was different, and our findings may not fully represent the stakeholder relationships and perceptions within a single event. It would be interesting to observe the caller-callee relationship within a single calling-out event and compare how the perspectives and perceptions differ. It was also noted by many participants that many callers were minors in their experience, which could have made them ineligible to participate in the current study. Future work utilizing data-driven analysis and large-scale modelling could give more quantitative insights into the overall phenomenon of online calling out.

Also, due to our selected method of recruitment, there may have been sampling bias in the process of recruiting our participants. We chose to use a public Tweet for recruiting participants as we understood the potential of Twitter's amplification networks for it to reach a larger network, but we recognize that the existing Twitter networks of the researchers may have influenced or limited the reach of the Tweet. However, from the final list of participants, only one was part of the researchers' Twitter follower networks, and all other participants were reached indirectly. We also note that there may have been selection bias as participation was voluntary, and may have favored participants more open to share their experience.

We purposefully did not collect rich background information about participants so as to reduce the participants' burdens on signing up for the study. Moreover, as most of our participants stayed anonymous on Twitter, they were cautious of opening up personal information to the researcher. Further research that deals with the impact of users' socioeconomic or educational background in calling out behaviors could be meaningful in understanding how findings may generalize to different populations.

Finally, as this study was conducted only on Korean Twitter users, the perception towards Twitter features and harassment might differ according to the cultural background of the users. We suggest that conducting a similar study at a larger or a more global scale, possibly extending to other social media services as well, would be beneficial to understanding the connotations behind the online harassment experience.

6 CONCLUSION

In this paper, we identify common perceptions and patterns surrounding the phenomenon of calling out on Twitter. While we accept the potential of calling out behavior in terms of a justice standpoint (e.g. public criticism of a real crime) or a communications perspective, we establish that it has potential to progress into harassment unless kept appropriately in check. We note that the distinction between callers and callees, or even harassers and victims are situational, and that in the open communication environment of Twitter, fragmentation and division of community norms could be a cause for conflict and sometimes harassment. We also identify several elements on Twitter considered to have high correlation with harassing behaviors, and offer design implications to prevent and mitigate harassment. We suggest designing for de-escalation and providing indirect bystander intervention methods as possible directions for future work.

ACKNOWLEDGMENTS

This work was supported by Center for Digital Humanities and Computational Social Sciences (KAIST), and by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-00075, Artificial Intelligence Graduate School Program(KAIST)). We thank the members of KIXLAB for their help, support and guidance throughout this research, as well as the reviewers for their help in improving the paper. Finally, we thank the interviewees who graciously shared their experiences with us, without which writing this paper would not have been possible.

REFERENCES

[1] [n.d.]. Change your subscription privacy settings - YouTube Help. https://support.google.com/youtube/answer/7280190?hl=en

[2] [n.d.]. Urban Dictionary: LRT. https://www.urbandictionary.com/define.php?term=LRT

[3] [n.d.]. Urban Dictionary: ratioed. https://www.urbandictionary.com/define.php?term=ratioed

[4] 2021. Giving People More Control on Instagram and Facebook. https://about.instagram.com/blog/announcements/giving-people-more-control

[5] 2021. Introducing new tools to protect our community from abuse | Instagram Blog. https://about.instagram.com/blog/announcements/introducing-new-tools-to-protect-our-community-from-abuse

[6] Robert Asen. 2000. Seeking the "counter" in counterpublics. *Communication theory* 10, 4 (2000), 424–446.

[7] Julian Ausserhofer and Axel Maireder. 2013. National politics on Twitter: Structures and topics of a networked public sphere. *Information, communication & society* 16, 3 (2013), 291–314.

[8] Rajesh Basak, Niloy Ganguly, Shamik Sural, and Soumya K. Ghosh. 2016. Look Before You Shame: A Study on Shaming Activities on Twitter. In *Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion*. ACM Press, Montréal, Québec, Canada, 11–12. https://doi.org/10.1145/2872518.2889414

[9] Rajesh Basak, Shamik Sural, Niloy Ganguly, and Soumya K. Ghosh. 2019. Online Public Shaming on Twitter: Detection, Analysis, and Mitigation. *IEEE Transactions on Computational Social Systems* 6, 2 (April 2019), 208–220.

https://doi.org/10.1109/TCSS.2019.2895734
[10] Michael S Bernstein, Eytan Bakshy, Moira Burke, and Brian Karrer. 2013. Quantifying the invisible audience in social networks. In *Proceedings of the SIGCHI conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 21–30.
[11] Elisa Jayne Bienenstock, Phillip Bonacich, and Melvin Oliver. 1990. The effect of network density and homogeneity on attitude polarization. *Social Networks* 12, 2 (1990), 153–172.
[12] Lindsay Blackwell, Tianying Chen, Sarita Schoenebeck, and Cliff Lampe. 2018. When Online Harassment Is Perceived as Justified. *Proceedings of the International AAAI Conference on Web and Social Media* 12, 1 (2018), 10.
[13] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW (Dec. 2017), 1–19. https://doi.org/10.1145/3134659
[14] Yarimar Bonilla and Jonathan Rosa. 2015. #Ferguson: Digital protest, hashtag ethnography, and the racial politics of social media in the United States. *American ethnologist* 42, 1 (2015), 4–17.
[15] Gwen Bouvier. 2020. Racist call-outs and cancel culture on Twitter: The limitations of the platform's ability to define issues of social justice. *Discourse, Context & Media* 38 (Dec. 2020), 100431. https://doi.org/10.1016/j.dcm.2020.100431
[16] Gwen Bouvier and David Machin. 2021. What gets lost in Twitter 'cancel culture' hashtags? Calling out racists reveals some limitations of social justice campaigns. *Discourse & Society* 32, 3 (May 2021), 307–327. https://doi.org/10.1177/0957926520977215
[17] danah boyd. 2008. Why youth (heart) social network sites: The role of networked publics in teenage social life. *YOUTH, IDENTITY, AND DIGITAL MEDIA, David Buckingham, ed., The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning, The MIT Press, Cambridge, MA* 2007-16, 1 (2008), 119–142.
[18] danah boyd, Scott Golder, and Gilad Lotan. 2010. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *2010 43rd Hawaii international conference on system sciences*. IEEE, 1–10.
[19] William J Brady, Julian A Wills, John T Jost, Joshua A Tucker, and Jay J Van Bavel. 2017. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences* 114, 28 (2017), 7313–7318.
[20] André Brock Jr. 2020. *Distributed Blackness: African American Cybercultures*. NYU Press.
[21] Nicholas Brody. 2021. Bystander Intervention in Cyberbullying and Online Harassment: The Role of Expectancy Violations. *International Journal of Communication* 15 (2021), 21.
[22] Nicholas Brody and Anita L Vangelisti. 2016. Bystander intervention in cyberbullying. *Communication Monographs* 83, 1 (2016), 94–119.
[23] Collin Gifford Brooke. 2000. Forgetting to be (post) human: Media and memory in a kairotic age. *JAC* (2000), 775–795.
[24] Amanda Burgess-Proctor, Justin W Patchin, and Sameer Hinduja. 2009. Cyberbullying and online harassment: Reconceptualizing the victimization of adolescent girls. *Female crime victims: Reality reconsidered* (2009), 153–175.
[25] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 1217–1230. https://doi.org/10.1145/2998181.2998213
[26] Elanor Colleoni, Alessandro Rozza, and Adam Arvidsson. 2014. Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of communication* 64, 2 (2014), 317–332.
[27] Robyn M Cooper and Warren J Blumenfeld. 2012. Responses to cyberbullying: A descriptive analysis of the frequency of and impact on LGBT and allied youth. *Journal of LGBT Youth* 9, 2 (2012), 153–177.
[28] Kate Crawford and Tarleton Gillespie. 2016. What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society* 18, 3 (2016), 410–428.
[29] Meredith D. Clark. 2020. DRAG THEM: A brief etymology of so-called "cancel culture". *Communication and the Public* 5, 3-4 (Sept. 2020), 88–92. https://doi.org/10.1177/2057047320961562
[30] Munmun De Choudhury, Shagun Jhaver, Benjamin Sugar, and Ingmar Weber. 2016. Social media participation in an activist movement for racial equality. In *Tenth International AAAI Conference on Web and Social Media*. 92–101.
[31] Francine Dehue, Catherine Bolman, and Trijntje Völlink. 2008. Cyberbullying: Youngsters' experiences and parental perception. *CyberPsychology & Behavior* 11, 2 (2008), 217–223.
[32] Ann DeSmet, Sara Bastiaensens, Katrien Van Cleemput, Karolien Poels, Heidi Vandebosch, Greet Cardon, and Ilse De Bourdeaudhuij. 2016. Deciding whether to look after them, to like it, or leave it: A multidimensional analysis of predictors of positive and negative bystander behavior in cyberbullying among adolescents. *Computers in Human Behavior* 57 (2016), 398–415.
[33] Ann DeSmet, Sara Bastiaensens, Katrien Van Cleemput, Karolien Poels, Heidi Vandebosch, and Ilse De Bourdeaudhuij. 2012. Mobilizing bystanders of cyberbullying: an exploratory study into behavioural determinants of defending the victim. *Annual review of cybertherapy and telemedicine* 10 (2012), 58–63.

[34] Dominic DiFranzo, Samuel Hardman Taylor, Franccesca Kazerooni, Olivia D Wherry, and Natalya N Bazarova. 2018. Upstanding by design: Bystander intervention in cyberbullying. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 1–12.

[35] Fernando Domínguez-Hernández, Lars Bonell, and Alejandro Martínez-González. 2018. A systematic literature review of factors that moderate bystanders' actions in cyberbullying. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* 12, 4 (2018), 19 pages.

[36] Casey Fiesler, Cliff Lampe, and Amy S Bruckman. 2016. Reality and perception of copyright terms of service for online content creation. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*. Association for Computing Machinery, New York, NY, USA, 1450–1461.

[37] Bianca Fileborn. 2017. Justice 2.0: Street harassment victims' use of social media and online activism as sites of informal justice. *British journal of criminology* 57, 6 (2017), 1482–1501.

[38] Kiran Garimella, Ingmar Weber, and Munmun De Choudhury. 2016. Quote rts on twitter: usage of the new feature for political discourse. In *Proceedings of the 8th ACM Conference on Web Science*. Association for Computing Machinery, New York, NY, USA, 200–204.

[39] Ysabel Gerrard. 2018. Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society* 20, 12 (2018), 4492–4511.

[40] Carly Gieseler. 2019. *The Voices of #MeToo: From Grassroots Activism to a Viral Roar*. Rowman & Littlefield.

[41] Alisdair A Gillespie. 2006. Cyber-bullying and harassment of teenagers: The legal response. *Journal of Social Welfare & Family Law* 28, 2 (2006), 123–136.

[42] David Theo Goldberg. 2015. *Are we all postracial yet?* John Wiley & Sons.

[43] Martin Grandjean. 2016. A social network analysis of Twitter: Mapping the digital humanities community. *Cogent Arts & Humanities* 3, 1 (2016), 1171458.

[44] Max Halupka. 2014. Clicktivism: A Systematic Heuristic. *Policy & Internet* 6, 2 (2014), 115–132. https://doi.org/10.1002/1944-2866.POI355

[45] Max Halupka. 2018. The legitimisation of clicktivism. *Australian Journal of Political Science* 53, 1 (Jan. 2018), 130–141. https://doi.org/10.1080/10361146.2017.1416586

[46] Michael A Hogg, John C Turner, and Barbara Davidson. 1990. Polarized norms and social frames of reference: A test of the self-categorization theory of group polarization. *Basic and Applied Social Psychology* 11, 1 (1990), 77–100.

[47] Jane Im, Jill Dimond, Melody Berton, Una Lee, Katherine Mustelier, Mark S. Ackerman, and Eric Gilbert. 2021. Yes: Affirmative Consent as a Theoretical Framework for Understanding and Imagining Social Platforms. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–18. https://doi.org/10.1145/3411764.3445778

[48] Shagun Jhaver, Larry Chan, and Amy Bruckman. 2018. The view from the other side: The border between controversial speech and harassment on Kotaku in Action. *First Monday* 23, 2 (2018), 41 pages.

[49] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online Harassment and Content Moderation: The Case of Blocklists. *ACM Trans. Comput.-Hum. Interact.* 25, 2 (April 2018), 1–33. https://doi.org/10.1145/3185593

[50] Lisa M Jones, Kimberly J Mitchell, and Heather A Turner. 2015. Victim reports of bystander reactions to in-person and online peer harassment: A national survey of adolescents. *Journal of youth and adolescence* 44, 12 (2015), 2308–2320.

[51] Antara Kashyap. 2021. Cancel Culture: Threat to Freedom of Expression or a Form of Accountability? https://www.news18.com/news/movies/cancel-culture-threat-to-freedom-of-expression-or-a-form-of-accountability-3611918.html

[52] George Kennedy, Andrew McCollough, Edward Dixon, Alexei Bastidas, John Ryan, Chris Loo, and Saurav Sahay. 2017. Technology solutions to combat online harassment. In *Proceedings of the first workshop on abusive language online*. 73–77.

[53] Eun-mee Kim and Jennifer Ihm. 2020. More than virality: Online sharing of controversial news with activated audience. *Journalism & Mass Communication Quarterly* 97, 1 (2020), 118–140.

[54] Hyunwoo Kim, Haesoo Kim, Kyung Je Jo, and Juho Kim. 2021. StarryThoughts: Facilitating Diverse Opinion Exploration on Social Issues. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–29.

[55] Jiyoun Kim. 2021. The Meaning of Numbers: Effect of Social Media Engagement Metrics in Risk Communication. *Communication Studies* 72, 2 (2021), 195–213.

[56] Mathias Klang and Nora Madison. 2018. Vigilantism or outrage: An exploration of policing social norms through social media. *Ethics for a Digital Age* 2 (2018), 151–165.

[57] Rob Kling, Ya-ching Lee, Al Teich, and Mark S Frankel. 1999. Assessing anonymous communication on the internet: Policy deliberations. *The Information Society* 15, 2 (1999), 79–90.

[58] Kate Klonick. 2015. A new taxonomy for online harms. *Boston University Law Review Annex* 95 (2015), 53–55.

[59] Travis Kriplean, Jonathan Morgan, Deen Freelon, Alan Borning, and Lance Bennett. 2012. Supporting reflective public thought with considerit. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*.

Association for Computing Machinery, New York, NY, USA, 265–274.
[60] Travis Kriplean, Michael Toomim, Jonathan Morgan, Alan Borning, and Amy Ko. 2012. Is this what you meant? Promoting listening on the web with reflect. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1559–1568.
[61] Michał Krzyżanowski and Per Ledin. 2017. Uncivility on the web: Populism in/and the borderline discourses of exclusion. *Journal of Language and Politics* 16, 4 (2017), 566–581.
[62] Ganaele Langlois, Greg Elmer, Fenwick McKelvey, and Zachary Devereaux. 2009. Networked Publics: The Double Articulation of Code and Politics on Facebook. *Canadian Journal of Communication* 34, 3 (2009).
[63] Colette Langos. 2012. Cyberbullying: The Challenge to Define. *Cyberpsychology, Behavior, and Social Networking* 15, 6 (June 2012), 285–289. https://doi.org/10.1089/cyber.2011.0588
[64] Yu-Hao Lee and Gary Hsieh. 2013. Does slacktivism hurt activism? The effects of moral balancing and consistency in online activism. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 811–820.
[65] Rebecca Lewis, Alice E Marwick, and William Clyde Partin. 2021. "We Dissect Stupidity and Respond to It": Response Videos and Networked Harassment on YouTube. *American Behavioral Scientist* 65, 5 (2021), 735–756.
[66] Hanlin Li, Disha Bora, Sagar Salvi, and Erin Brady. 2018. Slacktivists or Activists?: Identity Work in the Virtual Disability March. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–13. https://doi.org/10.1145/3173574.3173799
[67] Kwan Hui Lim and Amitava Datta. 2012. Finding twitter communities with common interests using following links of celebrities. In *Proceedings of the 3rd international workshop on Modeling social media*. 25–32.
[68] Lindsay Ellis. 2021. Mask Off. https://www.youtube.com/watch?v=C7aWz8q_IM4
[69] Zhe Liu and Ingmar Weber. 2014. Is Twitter a public sphere for online conflicts? A cross-ideological and cross-hierarchical look. In *International Conference on Social Informatics*. Springer, 336–347.
[70] Paul Benjamin Lowry, Jun Zhang, Chuang Wang, and Mikko Siponen. 2016. Why do adults engage in cyberbullying on social media? An integration of online disinhibition and deindividuation effects with the social structure and social learning model. *Information Systems Research* 27, 4 (2016), 962–986.
[71] Kaitlin Mahar, Amy X. Zhang, and David Karger. 2018. Squadbox: A Tool to Combat Email Harassment Using Friendsourced Moderation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–13. https://doi.org/10.1145/3173574.3174160
[72] Tolba Marwa, Ouadfel Salima, and Meshoul Souham. 2018. Deep learning for online harassment detection in tweets. In *2018 3rd International Conference on Pattern Analysis and Intelligent Systems (PAIS)*. IEEE, 1–5.
[73] Alice Marwick and danah boyd. 2014. 'It's just drama': Teen perspectives on conflict and aggression in a networked era. *Journal of youth studies* 17, 9 (2014), 1187–1204.
[74] Alice E Marwick. 2021. Morally motivated networked harassment as normative reinforcement. *Social Media+ Society* 7, 2 (2021), 20563051211021378.
[75] Alice E Marwick and danah boyd. 2011. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New media & society* 13, 1 (2011), 114–133.
[76] J Nathan Matias, Amy Johnson, Whitney Erin Boesel, Brian Keegan, Jaclyn Friedman, and Charlie DeTar. 2015. Reporting, reviewing, and responding to harassment on Twitter. *arXiv preprint arXiv:1505.03359* (2015), 63 pages.
[77] Viktor Mayer-Schönberger. 2011. *Delete: The virtue of forgetting in the digital age*. Princeton University Press.
[78] Anthony McCosker. 2015. Social Media Activism at the Margins: Managing Visibility, Voice and Vitality Affects. *Social Media + Society* 1, 2 (July 2015), 205630511560586. https://doi.org/10.1177/2056305115605860
[79] Kaitlynn Mendes, Jessica Ringrose, and Jessalynn Keller. 2018. #MeToo and the promise and pitfalls of challenging rape culture through digital feminist activism. *European Journal of Women's Studies* 25, 2 (May 2018), 236–246. https://doi.org/10.1177/1350506818765318
[80] Ersilia Menesini and Annalaura Nocentini. 2009. Cyberbullying Definition and Measurement: Some Critical Considerations. *Zeitschrift für Psychologie / Journal of Psychology* 217, 4 (Jan. 2009), 230–232. https://doi.org/10.1027/0044-3409.217.4.230
[81] Sanja Milivojevic and Alyce McGovern. 2014. The death of Jill Meagher: Crime and punishment on social media. *International journal for crime, justice and social democracy* 3, 3 (2014), 22–39.
[82] Logan Molyneux and Rachel R Mourão. 2019. Political journalists' normalization of Twitter: Interaction and new affordances. *Journalism Studies* 20, 2 (2019), 248–266.
[83] Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society* 20, 11 (2018), 4366–4383.
[84] Lisa Nakamura. 2015. The unwanted labour of social media: Women of colour call out culture as venture community management. *New Formations* 86, 86 (2015), 106–112.
[85] Eve Ng. 2020. No Grand Pronouncements Here...: Reflections on Cancel Culture and Digital Media Participation. *Television & New Media* 21, 6 (Sept. 2020), 621–627. https://doi.org/10.1177/1527476420918828

[86] Pippa Norris. 2021. Cancel culture: Myth or reality? *Political Studies* (2021), 00323217211037023.
[87] Fayika Farhat Nova, Michael Ann DeVito, Pratyasha Saha, Kazi Shohanur Rashid, Shashwata Roy Turzo, Sadia Afrin, and Shion Guha. 2021. " Facebook Promotes More Harassment" Social Media Ecosystem, Skill and Marginalized Hijra Identity in Bangladesh. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–35.
[88] Brian L Ott. 2017. The age of Twitter: Donald J. Trump and the politics of debasement. *Critical studies in media communication* 34, 1 (2017), 59–68.
[89] E. Pariser. 2011. *The Filter Bubble: What The Internet Is Hiding From You.* Penguin Books Limited.
[90] Jessica A Pater, Moon K Kim, Elizabeth D Mynatt, and Casey Fiesler. 2016. Characterizations of online harassment: Comparing policies across social media platforms. In *Proceedings of the 19th international conference on supporting group work*. 369–374.
[91] Inbal Peleg-Koriat and Carmit Klar-Chalamish. 2020. The #MeToo movement and restorative justice: exploring the views of the public. *Contemporary Justice Review* 23, 3 (July 2020), 239–260. https://doi.org/10.1080/10282580.2020.1783257
[92] Kususanto Prihadi, Yen Ling Hui, Melissa Chua, and Calvin KW Chang. 2019. Cyber-Victimization and Perceived Depression: Serial Mediation of Self-Esteem and Learned-Helplessness. *International Journal of Evaluation and Research in Education* 8, 4 (2019), 563–574.
[93] Hemant Purohit, Andrew Hampton, Valerie L Shalin, Amit P Sheth, John Flach, and Shreyansh Bhatt. 2013. What kind of #conversation is Twitter? Mining #psycholinguistic cues for emergency coordination. *Computers in Human Behavior* 29, 6 (2013), 2438–2447.
[94] Jon Ronson. 2015. How One Stupid Tweet Blew Up Justine Sacco's Life. *The New York Times* (Feb. 2015). https://www.nytimes.com/2015/02/15/magazine/how-one-stupid-tweet-ruined-justine-saccos-life.html
[95] Jon Ronson. 2016. *So You've Been Publicly Shamed.* Riverhead Books.
[96] Miia Sainio, René Veenstra, Gijs Huitsing, and Christina Salmivalli. 2011. Victims and their defenders: A dyadic approach. *International journal of behavioral development* 35, 2 (2011), 144–151.
[97] Michael Salter. 2013. Justice and revenge in online counter-publics: Emerging responses to sexual violence in the age of social media. *Crime, Media, Culture* 9, 3 (2013), 225–242.
[98] Amit M Schejter and Noam Tirosh. 2015. "Seek the meek, seek the just": Social media and social justice. *Telecommunications policy* 39, 9 (2015), 796–803.
[99] Sarita Schoenebeck, Oliver L Haimson, and Lisa Nakamura. 2021. Drawing from justice theories to support targets of online harassment. *New Media & Society* 23, 5 (May 2021), 1278–1300. https://doi.org/10.1177/1461444820913122
[100] Sarita Schoenebeck, Carol F. Scott, Emma Grace Hurley, Tammy Chang, and Ellen Selkie. 2021. Youth Trust in Social Media Companies and Expectations of Justice: Accountability and Repair After Online Harassment. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1 (April 2021), 1–18. https://doi.org/10.1145/3449076
[101] Martin EP Seligman. 1972. Learned helplessness. *Annual review of medicine* 23, 1 (1972), 407–412.
[102] Robert Slonje and Peter K Smith. 2008. Cyberbullying: Another main type of bullying? *Scandinavian journal of psychology* 49, 2 (2008), 147–154.
[103] Peter K Smith, Cristina Del Barrio, and Robert S Tokunaga. 2012. Definitions of Bullying and Cyberbullying: How Useful Are the Terms? In *Principles of cyberbullying research*. Routledge, 54–68.
[104] Dominic Spohr. 2017. Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *Business Information Review* 34, 3 (2017), 150–160.
[105] Maya Stewart and Ulrike Schultze. 2019. Producing solidarity in social media activism: The case of My Stealthy Freedom. *Information and Organization* 29, 3 (Sept. 2019), 100251. https://doi.org/10.1016/j.infoandorg.2019.04.003
[106] John Suler. 2004. The online disinhibition effect. *Cyberpsychology & behavior* 7, 3 (2004), 321–326.
[107] Abraham Ronel Martínez Teutle. 2010. Twitter: Network properties analysis. In *2010 20th International Conference on Electronics Communications and Computers (CONIELECOMP)*. 180–186. https://doi.org/10.1109/CONIELECOMP.2010.5440773
[108] Twitter Safety. 2021. Your Tweets = Your space. Now you can change who can reply to you even after you Tweet. https://t.co/3HFSjAotg7. https://twitter.com/TwitterSafety/status/1415025551773892608
[109] Twitter Support. 2020. Tweets about a Tweet add more to the conversation, so we've made them even easier to find. Retweets with comments are now called Quote Tweets and they've joined the Tweet detail view. Tap into a Tweet, then tap "Quote Tweets" to see them all in one place. https://t.co/kMqea6AC80. https://twitter.com/TwitterSupport/status/1300555325750292480
[110] Joseph Ching Velasco. 2020. You are Cancelled: Virtual Collective Consciousness and the Emergence of Cancel Culture as Ideological Purging. *rupkatha* 12, 5 (Oct. 2020), 7 pages. https://doi.org/10.21659/rupkatha.v12n5.rioc1s21n2
[111] Jessica Vitak, Kalyani Chadha, Linda Steiner, and Zahra Ashktorab. 2017. Identifying Women's Experiences With and Strategies for Mitigating Negative Effects of Online Harassment. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, Portland Oregon USA, 1231–1245. https:

//doi.org/10.1145/2998181.2998337
[112] Ariadne Vromen. 2017. Digital citizenship and political engagement. In *Digital citizenship and political engagement*. Springer, 9–49.
[113] Yu Wang, Jiebo Luo, Richard Niemi, Yuncheng Li, and Tianran Hu. 2016. Catching fire via "likes": Inferring topic preferences of Trump followers on twitter. In *Tenth International AAAI Conference on Web and Social Media*. 719–722.
[114] Michael Wenzel and Tyler G Okimoto. 2016. Retributive justice. In *Handbook of social justice theory and research*. Springer, 237–256.
[115] Randy Yee Man Wong, Christy MK Cheung, Bo Xiao, and Jason Bennett Thatcher. 2021. Standing up or standing by: Understanding bystanders' proactive reporting responses to social media harassment. *Information Systems Research* 32, 2 (2021), 561–581.
[116] Shuzhe Yang, Anabel Quan-Haase, and Kai Rannenberg. 2017. The changing public sphere on Twitter: Network structure, elites and topics of the #righttobeforgotten. *New media & society* 19, 12 (2017), 1983–2002.
[117] Sarita Yardi and danah boyd. 2010. Dynamic debates: An analysis of group polarization over time on twitter. *Bulletin of science, technology & society* 30, 5 (2010), 316–327.
[118] Amy X. Zhang and Scott Counts. 2015. *Modeling Ideology and Predicting Policy Change with Social Media: Case of Same-Sex Marriage*. Association for Computing Machinery, New York, NY, USA, 2603–2612. https://doi.org/10.1145/2702123.2702193

Received January 2022; revised April 2022; accepted May 2022