

석사학위논문
Master's Thesis

온라인상의 비판 현상에 대한 분석 및
사이버불링 유발 방지를 위한 플랫폼 디자인

An Observation of Online Call-out Culture:
Motivations, Repercussions, and Solutions to Online Harassment

2023

김해수 (金海秀 Kim, Haesoo)

한국과학기술원

Korea Advanced Institute of Science and Technology

석사학위논문

온라인상의 비판 현상에 대한 분석 및
사이버불링 유발 방지를 위한 플랫폼 디자인

2023

김해수

한국과학기술원

전산학부

온라인상의 비판 현상에 대한 분석 및 사이버불링 유발 방지를 위한 플랫폼 디자인

김 해 수

위 논문은 한국과학기술원 석사학위논문으로
학위논문 심사위원회의 심사를 통과하였음

2022년 12월 12일

심사위원장 장 정 우 

심사위원 김 주 호 

심사위원 차 미 영 

An Observation of Online Call-out Culture: Motivations, Repercussions, and Solutions to Online Harassment

Haesoo Kim

Major Advisor: Jeong-woo Jang

Co-Advisor: Juho Kim

A dissertation submitted to the faculty of
Korea Advanced Institute of Science and Technology in
partial fulfillment of the requirements for the degree of
Master of Science in Computer Science

Daejeon, Korea
December 12, 2022

Approved by



Jeong-woo Jang
Professor of Digital Humanities and Computational Social Sciences

The study was conducted in accordance with Code of Research Ethics¹.

¹ Declaration of Ethical Conduct in Research: I, as a graduate student of Korea Advanced Institute of Science and Technology, hereby declare that I have not committed any act that may damage the credibility of my research. This includes, but is not limited to, falsification, thesis written by someone else, distortion of research findings, and plagiarism. I confirm that my thesis contains honest conclusions based on my own careful research under the guidance of my advisor.

MCS

김해수. 온라인상의 비판 현상에 대한 분석 및 사이버불링 유발 방지를 위한 플랫폼 디자인. 전산학부 . 2023년. 60+iv 쪽. 지도교수: 장정우, 김주호. (영문 논문)

Haesoo Kim. An Observation of Online Call-out Culture: Motivations, Repercussions, and Solutions to Online Harassment. School of Computing . 2023. 60+iv pages. Advisor: Jeong-woo Jang, Juho Kim. (Text in English)

초 록

최근 온라인상에서는 타인에 대한 비판을 공개적인 장소에서 전시하는 현상이 더욱 잦아지고 있다. 이와 같은 현상은 자연스러운 소통의 일면일 수도 있으나, 개인에 대한 공격을 정당화하면서 온라인 괴롭힘, 혹은 사이버불링으로 발현할 가능성을 지닌다. 이와 같이 개인의 작은 잘못을 과도하게 처벌하고자 하는 문화가 발달할 경우, 온라인 공간이 공론장으로서 가지는 가치가 감소할 수 있다는 우려 또한 존재한다. 이와 같은 현상을 관찰하기 위해, 본 연구에서는 32명의 트위터 사용자를 대상으로 온라인상의 공개적인 비판 행위와 관련한 경험에 대한 인터뷰 연구를 진행했다. 이에 대한 결과로, 정당한 비판과 사이버불링을 구분짓는 데에는 비판 당사자가 대응할 수 있는지의 여부가 큰 영향을 끼친다는 것을 발견했다. 또한 사용자에 따라 비판 상황을 인지하는 방식이 달라지고, 이것은 사이버불링을 정의하는 데 있어 추가적인 영향을 끼친다는 것을 발견했다. 이와 같은 결과를 바탕으로 본 논문에서는 온라인 괴롭힘의 부정적인 영향을 축소하고, 비판 당사자가 효과적인 대응을 할 수 있도록 돕는 시스템인 Re:SPect 을 구상했다. 이에 더불어 본 논문에서는 온라인상에서 건설적인 대화와 토론을 하고 사이버불링을 방지할 수 있는 방법들에 대해 논의한다.

핵심 낱말 온라인 괴롭힘, 사이버불링, 소셜 미디어, 트위터, 소셜 미디어 디자인

Abstract

Calling out, a phenomenon where people publicly broadcast their critiques of someone to a larger audience, 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 experiences surrounding calling out on Twitter. We found that a key determining factor that distinguishes criticism from harassment was the subject's ability to respond to or engage with the criticism, and that different stakeholders hold different perspectives toward how online harassment is defined. Based on these findings, we explore design approaches that could reduce the negative effects of calling out and networked online harassment. Through an iterative design process, we introduce Re:SPect, a system designed to facilitate scalable responses from victims of networked online harassment. Finally, we discuss design implications for the platform in promoting healthy discourse while preventing toxic behavior on social media.

Keywords Online harassment, social media, Twitter, social media platform design

Contents

Contents	i
List of Tables	iii
List of Figures	iv
Chapter 1. Introduction	1
1.1 Position Statement	1
1.2 Thesis Contribution	2
Chapter 2. Background and Related Work	3
2.1 Calling out Behavior in Online Spaces	3
2.1.1 Performing Justice in Social Media	3
2.2 Challenges in Defining Online Harassment	4
2.3 Responding to Online Harassment	5
Chapter 3. An Investigation of Calling Out Behaviors in Twitter	6
3.1 Introduction	6
3.1.1 Research Context: Communication Features on Twitter	6
3.2 Methods	7
3.2.1 Participants	7
3.2.2 Interviews	9
3.3 Results	9
3.3.1 Patterns of Calling Out on Twitter	10
3.3.2 Why do People Call Out Others?	12
3.3.3 What Happens After Someone is Called Out?	15
3.3.4 How does the Twitter Community Assess Calling Outs?	18
3.3.5 How do Calling Outs Escalate to Harassment?	19
Chapter 4. Exploring System Designs to Mitigate Networked Online Harassment	23
4.1 Introduction	23
4.2 Design Iteration: Design Workshop	23
4.2.1 Methods	25
4.2.2 Workshop Materials	26
4.2.3 Participants	27
4.2.4 Workshop Results	27

4.2.5	Design Goals	29
4.3	Re:SPect	29
4.3.1	System Features	30
4.3.2	User Scenarios	33
4.4	User Evaluation	33
4.4.1	Methods	34
4.4.2	Results	34
Chapter 5.	Discussion	39
5.1	Implications for Discourse on Social Media	39
5.1.1	Limited Communicative Value of Calling out	39
5.1.2	Alienation of Callees Through Amplification	39
5.1.3	Impact on Willingness to Speak Up	40
5.2	Platform Dynamics in Calling Out and Harassment	40
5.2.1	Forming Distinct Sub-communities	41
5.2.2	Limitations of Response Measures	41
5.2.3	Amplification Features Promoting Harassment	42
5.2.4	Visible Engagement Metrics	42
5.3	Extending the Definition of Online Harassment	42
5.3.1	The Role of Context in Calling Out and Harassment	43
5.3.2	Unintentional Harassment	43
5.3.3	Interchangeability of Roles	43
5.3.4	Callees' Ability to Engage	44
5.4	Designing to Prevent Online Harassment	44
5.4.1	Employing an Experience-Centric Paradigm of Online Harassment	44
5.4.2	Designing for De-escalation	45
5.4.3	Providing Indirect Routes for Bystander Intervention	45
Chapter 6.	Conclusion	47
6.1	Generalizing Across Diverse Social Media Platforms	47
6.2	Limitations and Future Work	47
	Bibliography	49
	Acknowledgments	59
	Curriculum Vitae in Korean	60

List of Tables

3.1	Interview Participant Demographics.	8
3.2	Patterns and Motivations of Calling Out	12
3.3	Common response patterns of callees	16
4.1	Design Workshop Participant Demographics.	27
4.2	Re:SPect User Study Participant Demographics.	35

List of Figures

3.1	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.	10
4.1	(Left) Example of ‘Start Discussion’ feature embedded on Twitter. (Right) Example of a Tweet referencing and abstracted Tweet	24
4.2	Example of text-level abstraction. Text is ordered based on degree of abstraction	24
4.3	(Left) Example of an aggregated discussion thread. (Right) Users can search for discussion threads with the unique link generated for each Tweet	25
4.4	Workshop Stages and Protocol	26
4.5	Dashboard View of Re:SPect. (Left) Basic Dashboard View that shows the clustered responses. (Right) Detail View of each response cluster.	30
4.6	Manage Interactions Panel from the Dashboard View	31
4.7	Viewing Responses in Re:SPect based on the profile visibility conditions. (Left) When the viewer is outside of the profile visibility boundary. The viewer cannot access the original poster’s account information. (Right) When the viewer is within the profile visibility boundary.	31
4.8	Examples of post flags. They alert the viewer to the fact that there is additional context information that has been noted by the original post author. (Left) An ‘Additional Context Added’ flag is added to the callee’s post. (Right) A ‘Point Refuted’ flag is added to the caller’s post.	32

Chapter 1. Introduction

Since its conception, online social media has been a space for users to share their opinions and thoughts on various social issues. People communicate their interests on social media platforms, discuss controversial issues [1], and even participate in political discussions [2] - creating smaller topic networks and sub-communities in the process. As these topic networks are formed, social networks operate as a public sphere [2, 3] where various social issues are discussed through open communication [4]. While some have pointed out the limitations of such online public spheres in facilitating true democratic conversation [5, 6], there have been cases where such conversations extended past the online space and brought significant changes in the ‘real world’ as well [7, 8].

In some cases, these conversations happen through criticism - pointing out and raising awareness about issues that previously might not have been as visible. A particular method of criticism that has recently gained prominence is ‘calling out’: the public identification and criticism of individuals online [9, 10, 11]. Sometimes referred to as ‘cancel culture’ [12, 13], 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 [14]. 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 [15, 16]. 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 [17, 18]. In such cases, morally-motivated critiques toward the individual, while well-intended, could easily progress into online harassment, or more specifically, Networked Harassment [19, 20].

Networked harassment is defined as “online harassment against a target that is encouraged or instigated by members of an online network” [20]. Networked harassment is different from previous notions of bullying or harassment; because it functions primarily at scale, it does not map to traditional or legal models of harassment [21]. Furthermore, it may even be instigated unintentionally. For example, someone may benignly comment on a ‘problematic’ post, aiming to initiate conversation, but be interpreted as sealionin [19] or even encouraging harassment by exposing them to a larger networked audience [20]. In this thesis, we explore how online calling out and critical communication can be connected to networked harassment, as well as explore methods to alleviate the negative effects of calling out and networked online harassment.

1.1 Position Statement

We pause here to clarify the position of the author 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. For example, in the context of Twitter, the number of followers - or supporters - can create privileges and power structures independent of their position within society. Following the concept of intersectionality, we recognize that multiple forms of inequalities may combine or overlap to create unique experiences that may not be fully represented in the current work.

As such, power dynamics and marginalization are not absolute concepts, and might differ according to the specific situation that the individual is facing. We do not claim to speak for the entirety of experiences surrounding online calling out and harassment, but rather provide a lens into the individual experiences that represent how such relations and dynamics might present themselves.

1.2 Thesis Contribution

This thesis makes the following major contributions:

- A descriptive model of online calling out behavior, including common stages and patterns based on diverse user experiences
- Insights for what may distinguish between online harassment from ‘valid’ criticism in the perspective of Twitter users
- Design implications for online social media platforms in preventing online harassment while simultaneously encouraging healthy discourse

Chapter 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. Finally, we discuss what efforts have been made to combat online harassment in social media platforms.

2.1 Calling out Behavior in Online Spaces

Public criticism behaviors in social media have been referred to in various ways, including ‘cancelling’ [22, 13], public shaming [18] and calling out [16]. 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 [23]. However, these terms were often used with negative connotations, implying that it has become a trivial habit of the public [13], or even a case of mob mentality where users would simply ‘attack’ people [23].

As many calling out cases happen on Twitter [22, 17, 24], 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 [25, 26]. As calling outs became prevalent, casual terms such as “Twitter’s villain of the day” [27] have also emerged, implying the commonality of calling outs. These limitations have been attributed to the relative lack of effort involved in tweeting [17], the high speed with which text is disseminated [28], as well as the lack of nuance in the limited space [14]. Bouvier observed that tweets using a ‘cancel culture’ hashtag would often represent racism as a personal, homogeneous trait [17], 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 [29, 25].

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 [18]. 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 [17].

2.1.1 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 [30]. 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 chan-

nels [31]. This emphasizes the role of the internet to operate as a counter-public space [32, 31], challenging existing communicative hegemonies through consciousness-raising [33] and redemocratizing public conversation [34, 35, 31].

Calling out behaviors have been used as an attempt at restoring justice where criminal justice laws could not perpetrate [36]. As youth are less likely to trust social media companies or legal authorities to achieve fair resolutions in social media disputes [37], they instead turn to more personal modes of intervention such as criticism [38] or a public demand for an apology [39]. 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 [15, 7].

On a society-wide scale, social media has been considered a valid platform for performing identity, solidarity, and activism, especially for minority groups [40, 41, 42]. 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 [43]. In particular, micro-political activities [44] have been referred to as ‘slacktivism’ [43] or ‘clicktivism’ [45], 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 [45, 46].

Finally, users may attempt to take matters into their own hands. While ‘cancel culture’ focuses on high-profile individuals such as politicians and celebrities [24], everyday individuals are also subject to such scrutiny when they are perceived to have done something wrong [47, 14]. This can be observed in a retributive justice standpoint, which suggests that individuals receive a proportional, ‘deserved’ punishment for their actions [48]. Blackwell et al. explored how retributive approaches are perceived by social media users in response to a perceived transgression [38]. Marwick introduced the concept of Morally Motivated Networked Harassment (MMNH) where people utilize networked harassment to reinforce social and moral norms [20]. Here, people use calling out behaviors as a form of social shaming, upholding social norms by publicly humiliating the callees [49]. 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. [47] 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.

2.2 Challenges in Defining Online Harassment

Networked online spaces are fundamentally different from offline, unmediated spaces in terms of its persistence, searchability, replicability, and the invisible audiences [50]. 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 [51], and some have noted that it may cause more psychological damage than offline bullying [52]. Anonymity also has a significant impact on online harassment, as it can foster disinhibition and deindividuation within users [53], reducing their sense of responsibility [54, 55] and magnifying deviant behaviors [56].

Traditional definitions of bullying include elements such as repetition of messages, power differential between the perpetrator and victim, intent to harm, and aggression [21]. 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 [57, 58] as well as distributed to a larger potential audience [59]. 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 [21, 60, 61].

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 [19]. 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’ [19, 38, 62]). 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 [63]. Others have voiced concerns about the desensitization due to the prevalent harassing behaviors in online spaces [64].

2.3 Responding to Online Harassment

A significant body of work in HCI and social computing focus on methods of preventing or mitigating the effects of harassment. Most social media platforms adopt some form of content moderation to protect users against abusive behavior [65], but platforms usually do not have a clear definition of what constitutes abuse [66], nor are they well-communicated to their users [67, 68]. Moreover, as online content moderation usually focuses on punishing the offender [66], less attention has been made to address the impact on the targeted user [69]. 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 [39].

A significant body of research in NLP has focused on automatic detection of harassment, building datasets of harassing messages [70, 71] or detection models [72, 73, 74]. Other work, such as Recast aims to reduce harassment through detecting toxic language and intention, discouraging users from posting harmful messages [75]. However, while automated moderation could help scalable anti-harassment interventions, there are limitations to automated detection as methods of harassment continually evolve with the development of social technologies [76, 66].

When systematic efforts fall short, users have also leveraged community efforts to combat harassment. Community-created collaborative blocklists, such as BlockTogether and Good Game Auto Blocker are an example of how communities may come together to protect themselves from potential harassers [19]. Other platforms such as Heartmob [63] have provided safe spaces for victims to share their experiences, facilitating recovery and emotional support. Squadbox utilizes a user’s friend groups as a moderation tool, allowing for more personalized and intimate methods of protecting someone from harassment [60].

In addition to simply preventing harassment, social computing research has also explored how platforms might facilitate healthy, non-toxic discussions. Systems such as ConsiderIt [77] and Reflect [78] experimented encouraging users to consider differing viewpoints in a civil manner. Nelimarkka et al. suggested design recommendations on how to decrease polarization and facilitate discussion in social spaces [79]. Kim et al. explore the possibility of priming users to contextual information about someone’s post [80]. We borrow from such previous insights on reducing online hostility and facilitating discussion to design our own solution that can facilitate open discussion online, while still preventing the negative effects and of potential harassment.

Chapter 3. An Investigation of Calling Out Behaviors in Twitter

3.1 Introduction

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 [81] 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.

3.1.1 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 [82]; retweets (RT), where users directly repost messages posted by others [83]; quote-tweets (QT), where users are able to directly repost others’ tweets while adding their own comment as a new tweet [84]; and finally replies, a commenting format that adds and displays the reply in thread format from the original tweet [85].

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’ [86]. 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 [85], and direct messages (DMs), private messages accessible to only the sender and receiver.

3.2 Methods

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 [87], 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 different stakeholder groups perceive or evaluate calling outs differently?

RQ4. How do calling outs escalate into online harassment?

3.2.1 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 3.1](#). IDs indicate the primary category of participant, bystanders(**B**), callees(**E**), and callers(**R**). 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 note that this does not constrict the experience of each participant as many participants had experiences across multiple calling out incidents and categories. 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. 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

Table 3.1: Interview Participant Demographics.

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	

caller. We attempted to balance out the overall variety of experiences through increasing the number of caller participants.

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*) [88].

3.2.2 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.

3.3 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.

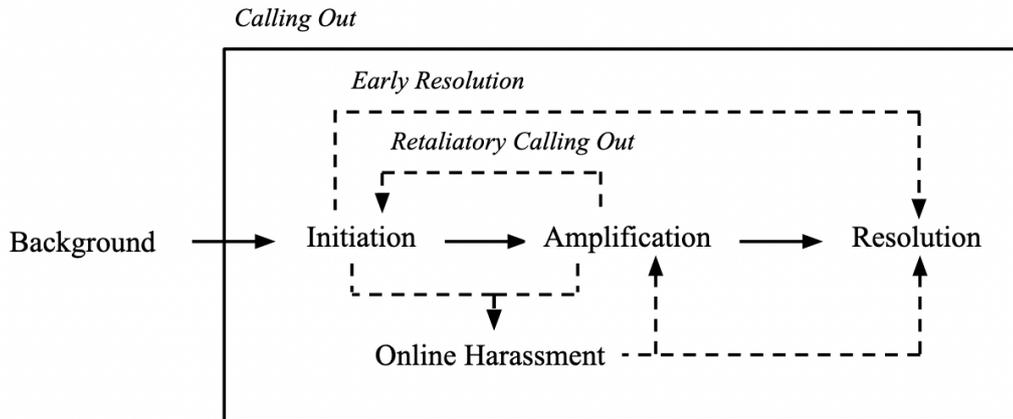


Figure 3.1: Lifecycle diagram of a calling out. Solid line arrows denote the general transition between phases, and dotted arrows denote notable deviations from the central lifecycle.

3.3.1 Patterns of 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.

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 3.1.

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 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.

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 3.2. 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

Table 3.2: Patterns and Motivations of Calling 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

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*).

3.3.2 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.

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.

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' [89], 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

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

3.3.3 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.

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

Table 3.3: 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).

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

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 the situation, there were several common approaches that callees would take. This is organized in Table 3.3. 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.

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 mute¹ those topics. - R4

3.3.4 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.

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

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

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

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.

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 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

3.3.5 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.

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.

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

Chapter 4. Exploring System Designs to Mitigate Networked Online Harassment

4.1 Introduction

Based on the findings from the previous study, we explore methods of how to mitigate the negative effects of networked harassment and how to facilitate constructive, healthy communication online. In particular, we focus on the perspective of victims of networked online harassment - what are their needs, and how can social media platforms support them.

While there has been a significant body of research surrounding how to respond to online harassment, there has not been much work on how to provide victims of harassment with the ability to resist. Much work has been done on preventing toxicity detection [72], where it can be used to moderate content on platforms [90, 73, 74, 70], or discourage individuals from posting potentially harmful content [75, 91]. Promoting bystander support has also been noted as a method to prevent online harassment [92, 93, 94]. However, there has been a relative lack of scholarship exploring active response measures that victims of harassment can take. Squadbox [60] applies friendsourced moderation as a way for individuals to exert more personalized, granular control, while systems such as Heartmob [63] provide support systems for individuals who have experienced harassment. Despite this, systematic support that focuses on what the individual can do is currently lacking.

Feminist and Queer scholarship has previously introduced the concept of *affirmative consent* - commonly characterized by the mantra “yes means yes”. Affirmative consent is a framework that argues that one must ask for - and earn - enthusiastic approval before interacting with another person [95]. Borrowing from feminist and queer ideologies, Im et al. have previously discussed the potential of applying such a framework to the design of social media systems [62]. In this perspective, mass online harassment is a violation of the victim’s consent: they are forced into a situation where they are being criticized at a scale they did not anticipate, and are not able to respond to effectively. Thus, we argue that by allowing harassment victims to exert their consent, we allow them to reclaim their agency and provide the ability to resist online harassment, while simultaneously protecting them against extreme harms.

In light of this, we explore the possibility of alleviating the issue of networked harassment by designing systems that allow users to maintain their agency in communication. In particular, we observe the potential of preventing users from being pulled into conversations without their consent, while providing scalable response measures in open online communication. To achieve this, we designed Re:SPect, a system that allows users to have granular, specific, and scalable controls over the discussion happening surrounding their posts.

4.2 Design Iteration: Design Workshop

We designed an initial version of the system with the goal of creating a social media platform that promotes consentful communication and scalable, practical responses to online harassment. We first designed our system with the concept of ‘separating’ the post, and the subsequent reactions to the post, from the person who posted it. Our approach to this was to provide a layer of abstraction between the

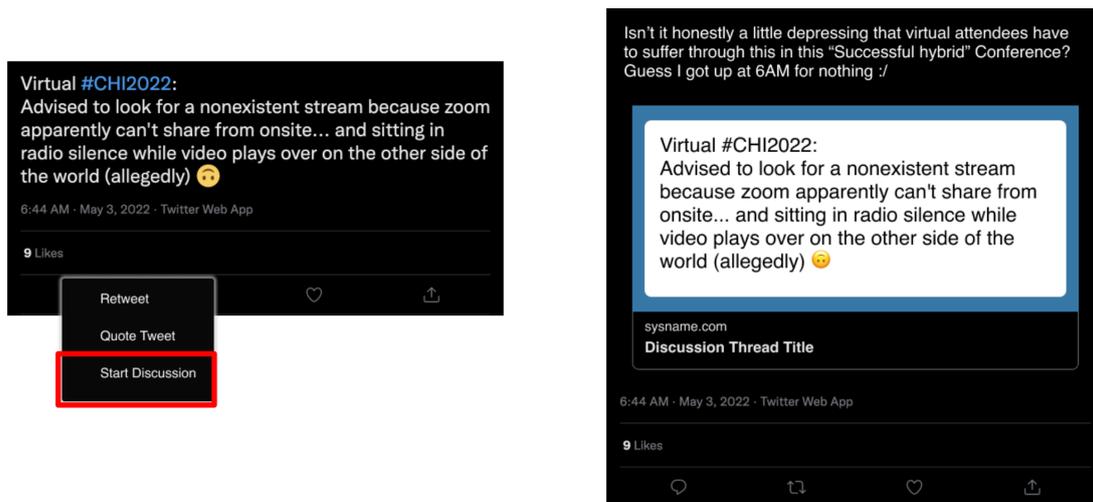


Figure 4.1: (Left) Example of ‘Start Discussion’ feature embedded on Twitter. (Right) Example of a Tweet referencing and abstracted Tweet

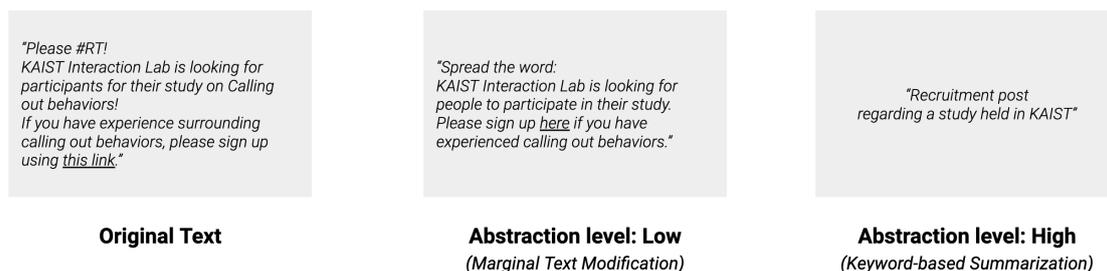


Figure 4.2: Example of text-level abstraction. Text is ordered based on degree of abstraction

person and the post, through a mediator system that hosts the discussion, independent from the original post.

Our initial concept was to implement our system as a browser extension that expands the functions of Twitter. When using the system, if a user (Caller) notices a Tweet, they are provided with the option, ‘Start Discussion’, in addition to the original Retweet and Quote-Tweet functions (Figure 4.1, Left). When a user selects this option, this would create an instance of the Tweet in our system, represented by the Tweet’s unique hash value. The caller is given the ability to add their opinion and post it as they would with the Quote-Tweet function. When posted, the Caller’s Tweet will represent the original tweet as an abstracted version, which leads to the mediated version uploaded in a separate database (Figure 4.1, Right). The mediated version does not save any identifying information pertaining to the original poster, such as profile information.

Every time a caller user makes a comment using the ‘Start Discussion’ feature, the system aggregates the individual discussion nodes to the original Tweet (Figure 4.3, Left). Users can access the aggregated discussion through each individual discussion thread, or by searching up the Tweet in the mediator

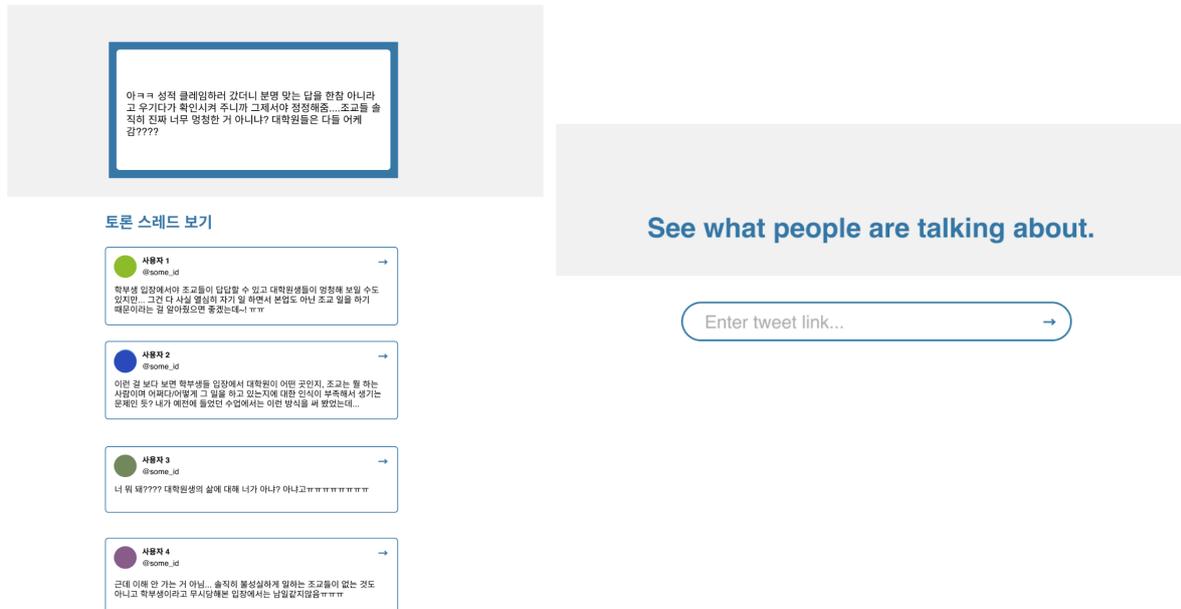


Figure 4.3: (Left) Example of an aggregated discussion thread. (Right) Users can search for discussion threads with the unique link generated for each Tweet

system. Thus, users can reach the discussion thread through instances of the subthread, or by searching with the original Tweet (Figure 4.3, Right), but the original Tweet or poster cannot be traced back from the discussion thread. To prevent searching for the post text content to find the user, we also implemented a further layer of abstraction, where the text will be obfuscated so that malicious users cannot trace it back to the original poster’s account and harass them. Examples of such text-level abstraction is depicted in Figure 4.2. The original poster (Callee; ‘Owner’ of the Tweet) also retains a level of control over the abstracted Tweet content, including the ability to delete the post and the subsequent discussion, as well as the level of text abstraction.

4.2.1 Methods

To improve the design and ensure that we create a system that accurately represents the users’ needs, we conducted a participatory design workshop with Twitter users to gather feedback on our initial design. We aimed to verify our design concepts, collect potential use case scenarios, and get a better understanding of how to prevent networked harassment.

The workshop consisted of 3 sections. The detailed workshop proceedings are depicted in Figure 4.4. First, we asked the participants to share their experiences and opinions surrounding public criticism and online harassment. Following that, we introduced our system concept and workflow, as well as potential use case scenarios. Participants were asked to suggest alternative features and new functions that could be added to the system, and imagine potential use cases and scenarios that might utilize our system. Finally, we moved on to a group interview session where the participants provided feedback on the approach and system design, and discussed what elements should be considered when building social media systems that combat online harassment. We then performed qualitative coding based on the workshop transcripts. Two researchers individually developed codes from the initial transcripts, which

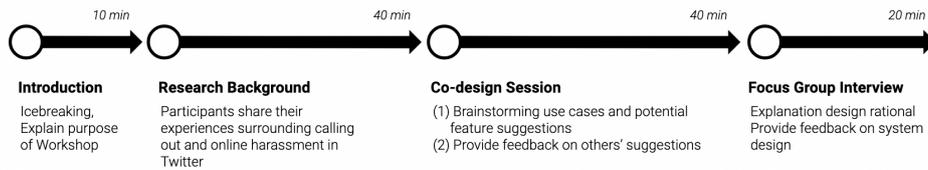


Figure 4.4: Workshop Stages and Protocol

were then combined to create a new codebook. The first author then used this codebook to re-code the transcripts.

4.2.2 Workshop Materials

To illustrate the motivations and mechanics of our system, we provided the users with screenshots of our system prototype as illustrated above, along with a set of potential use case scenarios for the system. We focused on two perspectives when building the scenario. The first scenario was based on the perspective of the potential harasser - who is invested in the topic being discussed as a result of the OP's post, but does not want to potentially participate in calling out or harassment by publicly responding to the post. The second scenario focused on the perspective of the victim, and how they would respond to being called out or harassed.

As the workshop was conducted in Korean, we provide a translated version of the suggested scenarios below.

Scenario A: Caller Perspective

User A is browsing Twitter when they discovers another user (User B) displaying behavior that they do not agree with. User A would like to criticize user B's behavior, but they know that their disagreement with user B's behavior is personal, and not necessarily ethically condemnable. User A is concerned about the possibility that their public criticism may introduce user B in harassment. Thus, user A utilizes our system to create a discussion thread based on an abstracted version of user B's Tweet, and adds their perspective on the discussion thread. User A is thus able to comment on the subject without directly interacting with or engaging user B. User C, a user who follows user A, discovers the discussion thread that user A posted, but they are not able to re-trace user B's account or user B's original post.

Scenario B: Callee Perspective

When user B logs onto Twitter, they get a notification from the system stating that a discussion has started based on their original Tweet. User B checks their notifications to access this discussion. As the 'owner' of the Tweet, user B retains the ability to control how their text is expressed or how it is repurposed through abstraction. While browsing through the options, user B notes that the experience that they Tweeted was a specific and personal experience, and thus other users might be able to specify them based on it. User B wants to make sure that other people cannot reach them based on the discussion thread, so they edit the level of abstraction to be more abstract. This hides the original text of the post and replaces it with a string of keywords that the original Tweet referred to. User B is also given additional controls with regards to the visibility settings, and even deleting the entire discussion thread.

Table 4.1: Design Workshop Participant Demographics.

ID	Gender	Group	Age	# of Accounts	Calling Out Experience		
					Caller	Callee	Bystander
W1	F	Group 1	21	1	O	O	O
W2	F	Group 1	24	4	O	O	O
W3	Other	Group 1	23	2	O	O	O
W4	F	Group 1	25	1	O	O	O
W5	F	Group 2	30	1	O	O	O
W6	F	Group 2	21	5	O	O	O
W7	Prefer not to say	Group 2	22	5	O	O	O
W8	F	Group 3	36	2		O	O
W9	F	Group 3	21	3		O	O
W10	F	Group 3	27	1		O	
W11	F	Group 3	40	4	O	O	O

4.2.3 Participants

Through a public Twitter post, we recruited 11 Twitter users who had experience surrounding online harassment, either as victims, perpetrators or bystanders. Participants were divided into groups of 3 to 4 people and participated in a two-hour design workshop session through a Zoom video call.

To ensure the participants’ safety and comfort, all participants communicated through pseudonyms that were assigned to them by the researchers, and communicated through audio only to prevent potentially identifying themselves. They were also advised to defer from disclosing sensitive or personally identifiable information. Participant ages ranged from 21 to 40 with an average of 26.36. The participant demographics are detailed in Table 4.1.

4.2.4 Workshop Results

Experience with Networked Harassment

Twitter’s platform design encourages harassment at scale Many participants shared the sentiment that bullying and harassment was perpetuated by the nature and design of Twitter as a social media platform (W7, W10, W11). W11 emphasized this by saying: “Twitter is a platform made perfect for bullying.” and that “users react strongly to differences in opinions [on Twitter].” Reasons for this was attributed to the word count limits that hinders nuanced communication (W11), the fact that amplification features would alert new users to already overheated conflicts (W7).

Some participants also generalized the phenomenon to social media platforms, W7 for example stating “I think fights occur very easily on social media.” W10 further articulated this, claiming that this is “Because anyone can participate in the conversation, it is also easier for unnecessary opinions to join the conversation, incurring fights more easily”. W7 also noted that “Everyone seeing everyone’s posts is theoretically beneficial, but on Twitter it becomes harmful.”, noting the potential harm that focusing on amplification features may bring.

Calling out as an Exertion of Power The majority of our workshop participants also stated or agreed with the belief that criticism becomes harassment when the critical opinion overwhelms the

supportive opinion. In other words, it was a game of numbers and power dynamics. Such power imbalances were stated to be based on the follower counts of the user. For example, W11 stated that “Followers are power.”, and described a situation where a callee was outnumbered by the caller group and had to avoid conflict despite not being in the wrong. Some participants recounted experiences where the user with a significantly larger following leveraged their users in targeting others. Others expressed fear or reluctance in criticizing or correcting other users if they had a much larger Twitter following.

W9 noted that some users also use calling out for ‘clout’, stating that “Some people criticize purely in order to gain more followers rather than to correct people or for moral reasons.” Therefore, calling out and the criticism that came from it was not necessarily seen as always justified. W2 also mentioned a similar point, pointing out that “Online harassment occurs when the user wants to feel safe among the masses.” She also added that “People want to know that their opinion is in the majority.” These comments imply that some Twitter users thought of online calling out as sometimes being more of a mob mentality behavior rather than stemming from actual criticism.

Design Implications for Preventing Harassment

Content should accurately reflect the context and intent of the original poster One of the major comments from the participants was that the abstracted content should fully and accurately transfer the original poster’s intent, tone, and content (W6, W10, W4). In response to the concept of ‘abstraction’ suggested in the design prototype, the participants raised concerns that changing the text or the content of the post can cause further misunderstandings or even worsen harassment. This was noted especially in relation to the word count limit on Twitter - the limited space means that diverse nuances and contexts are packed into a small amount of text, which could be hard to represent when the content is altered.

Participants also noted that users should retain agency over how their post is expressed and how others might perceive it, while preventing the potential of misuse of such features. For example, features allowing users to be able to edit the post was generally perceived negatively. Participants were concerned about the possibility that malicious users will post harmful content, edit the post, and then claim that the criticism that they are receiving is harassment. However, they were enthusiastic about the idea of providing methods to add context after a post was made, such as providing more visibility to added context. W1 mentioned that “Emphasizing edit tweets could be a good option.” and W3 said that they thought “it might be better to allow editing in the thread rather than in the post.”

Specific and granular control of notification settings In relation to the above point, participants voiced that the original poster should have full agency of the conversation. Participants noted that “The OP knows what parts induce stress for them. (W7)”, and that “What a person writes is part of how they express their identity. (W5)”. An example of this was granular notification settings. One of the major factors that distinguished criticism from harassment, as mentioned by the participants, was scale, which reflected the results from our previous study. Participants noted that part of this was due to the fact that the large influx of notifications caused social pressure, as well as a feeling of helplessness. Thus, many participants suggested that the system provide specific controls for notifications (W5, W10).

Prevent amplification and reduce transmissibility of posts Participants also noted that it is important to reduce the spread of the post to ensure that the negative effects of calling out and harassment could be mitigated (W2, W8, W10). This aligns with the findings from the previous study stating that the

scale of the calling out is a large determining factor of online harassment. W10 articulated this by saying: “I think the problem is receiving attention that you wouldn’t in real life.” Thus, the incomparable scale of the online response, in comparison to offline responses that the users are more accustomed to, would cause a large emotional reaction. In response to this, some participants suggested measures to allow the original poster to nip the harassment at the first signs with preventative measures. In particular, W2 said that “It’s important to cut off the interaction [with the harasser].” She further suggested that “If the amount of feedback increases rapidly, it might be good to briefly stop the interaction. I think it would be good if there was some kind of locking mechanism.” W8 also mentioned that ‘locking the spread’ of posts would be beneficial.

Emphasize responses encouraging constructive discussion Some participants noted that designs to reduce harassment would not necessarily be able to stop those who have specific intention to harass, as malicious users will eventually find a way (W5, W10, W11). Others also pointed out that simply cutting of interactions was not a healthy reaction, as it could cause echo chambers and people not being aware of their mistakes of constructive criticism . Therefore, one suggestion was to increase visibility or emphasize responses that encouraged constructive discussion, while reducing visibility of repetitive, aggressive, or harassing responses.

4.2.5 Design Goals

Based on these insights from the design workshop, we identified three design goals for designing a social media platform that protects users from the negative effects of networked harassment and calling out. These design goals each address a key need identified from Twitter users, which are: the ability to distance themselves from the conversation (**D1**), and being able to comprehend (**D2**) and respond to (**D3**) replies at scale.

D1. Protect victims of harassment from harassers by separating them from potential callees and allowing them to establish a safe distance from open audience spaces. (*Protect*)

D2. Provide a succinct, digestible summary of user comments to allow users to efficiently and accurately comprehend the content and state of the discussion, especially at a large scale (*Summarize*)

D3. Provide practical and scalable response measures to victims of large-scale harassment campaigns or calling outs (*Response*)

4.3 Re:SPect

Based on our design goals, we revised our original concept to design *Re:SPect*, a social media system that provides active and scalable response measures for victims of networked online harassment and calling out. In Re:SPect, we focused on the perspective of protecting the victim, as well as providing practical response measures for them to respond against online harassment. Furthermore, we aimed to suggest novel paradigms in ways to support victims of online networked harassment. In this section, we describe the features of Re:SPect, as well as potential user scenarios where Re:SPect can be used to support victims of online harassment.

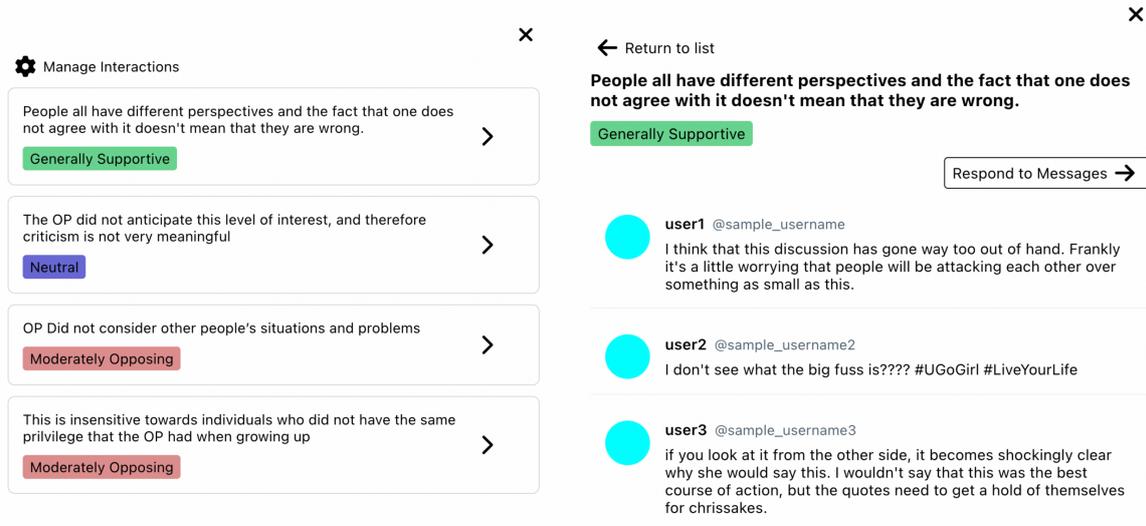


Figure 4.5: Dashboard View of Re:SPect. (Left) Basic Dashboard View that shows the clustered responses. (Right) Detail View of each response cluster.

4.3.1 System Features

The basic features of Re:SPect were built to emulate the design of Twitter, and similar social media platforms such as Mastodon¹. This was primarily due to the fact that we recruited Twitter users as our target users for evaluation. We wanted to minimize the level of unfamiliarity from our users as novel interface features might have a play in how users perceive the system or the proposed situation of networked harassment. While networked harassment can happen on platforms other than Twitter, the prevalence of networked harassment on Twitter as evidenced by our previous study, as well as the platform’s tendency to promote post amplification as one of their central features, led to our decision to focus on Twitter as the target platform.

When a user views their post, they are provided with the option to ‘See Overview of Discussion’. Clicking on this button leads them to the dashboard interface that organizes the information about the post as well as its responses. The dashboard view is shown in Figure 4.5.

Control who can See or Interact with the Post

In the dashboard view, the owner of the post can access the ‘Manage Interactions’ tab to control who can see or interact with their post (Fig 4.6). This controls how the post is displayed to others in three levels - the visibility of the post, the visibility of the user profile, and who can interact with the post. The levels are organized in descending order of exclusivity, meaning that controlling the post at the previous level automatically includes control at the subsequent, lower levels. When a user is outside of the distance conditions set by the post owner, they are unable to access the corresponding information. For example, if a user is outside of the profile visibility boundary, but within the post visibility boundary, they will see an anonymized post where they cannot trace the post back to the original poster (Fig 4.7).

Each setting has three standard options: (i) **Everyone**, meaning that everyone on the social network can access the post, (ii) **Followers-only**, where only the immediate follower network of the user can

¹<https://joinmastodon.org/>

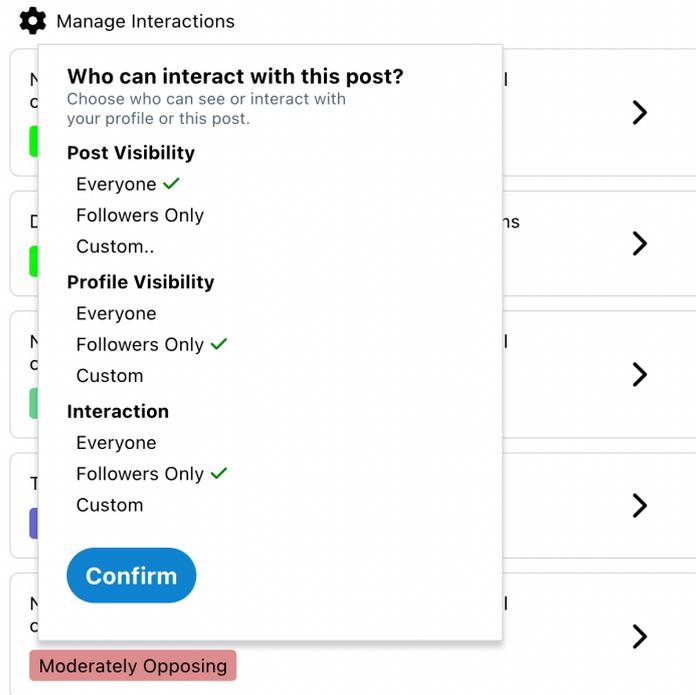


Figure 4.6: Manage Interactions Panel from the Dashboard View



Figure 4.7: Viewing Responses in Re:SPeCt based on the profile visibility conditions. (Left) When the viewer is outside of the profile visibility boundary. The viewer cannot access the original poster's account information. (Right) When the viewer is within the profile visibility boundary.

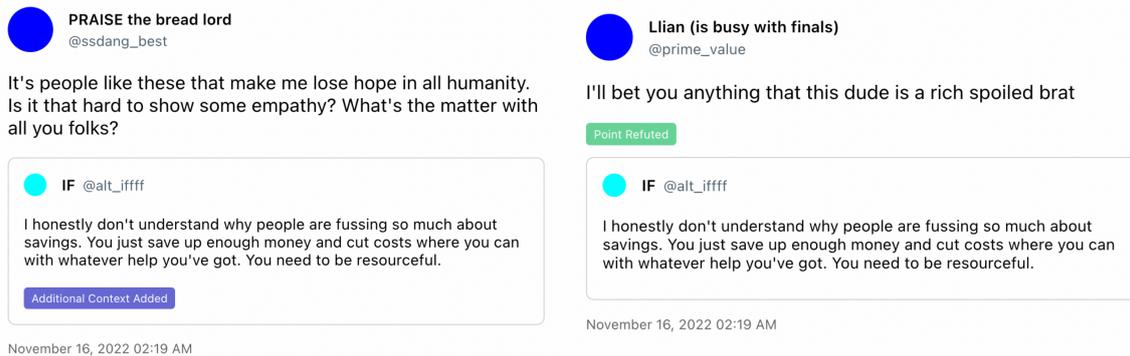


Figure 4.8: Examples of post flags. They alert the viewer to the fact that there is additional context information that has been noted by the original post author. (Left) An ‘Additional Context Added’ flag is added to the callee’s post. (Right) A ‘Point Refuted’ flag is added to the caller’s post.

access the post, and finally (iii) **Custom**, where the original poster (OP) can provide the distance to which their post can reach. For example, if the OP wants their post to be accessed by only those who follow the OP’s followers, the custom network distance would be set to 2. The distance number is determined according to the calculated network distance from the original poster.

Summarization of Responses

Re:SPect performs topic clustering on the responses to provide a digestible and summarized view of the responses on the original post. This provides users with the ability to objectively analyze the overall opinion distribution with regard to their post. This also has the benefit of filtering out malicious or aggressive comments, such that the user doesn’t have to continuously be introduced to each individual comment. This reduces the potential negative emotional repercussions of reading through each comment individually. Each topic cluster is then represented by its central argument, generated through text summarization of the individual response posts that are in the cluster. The topic cluster view presents the central argument, the general sentiment distribution of posts within the cluster. By clicking on the cluster object, the user is able to see what specific comments are in each cluster.

Responding to Comments at Scale

Users can also choose to respond to a specific cluster of comments to respond to. In the Response Cluster Tab, users can select the type of response measure to take. The system automatically mass-applies the selected measure to all the responses in the cluster. Users can also specify which specific posts within the cluster they want to respond to. This supports traditional response measures such as blocking the users, reporting them or posting individual replies, as well as a flagging mechanism that allows the original poster to add additional context to the original post while controlling how they are perceived by the networked audience.

The user may choose to flag either their original post or the replies to it. There are two types of flags applicable to the posts: ‘Additional Context Added’ and ‘Point Refuted’ (Figure 4.8). The former is applied to the original post, when the OP appends more contextual information to the original post that can help understand the full picture. The latter, on the other hand, is added to the response post,

when the OP wishes to express that they have provided a rebuttal to the critique made in the response. Both of these flags are designed to provide additional context that may have been lost due to the word length limit of platforms such as Twitter, as well as those that are lost when only individual posts gain attention instead of the full conversational context. Through these interface elements, users can alert the networked audience that the current limited perspective is not all that there is to the conversation, providing them with the ability to control others' perceptions about themselves.

4.3.2 User Scenarios

Callee Scenario

Julia, a climate activist, makes a post on social media criticizing a popular celebrity's actions that had a negative environmental impact. Her intention was to share it with her network of friends who are climate activists themselves. However, fans of the celebrity soon discover Julia's post and begin mass-posting harassing responses to her post. Overwhelmed by the number of responses, Julia enters the 'Discussion Overview' dashboard to see a summarized view of the responses. She sees that there are a fair amount of supportive comments who agree with the point she made, while the large number of malicious responses have been combined into one cluster. She selects the cluster, and proceeds to use Re:SPect to mass-block everyone who has commented rude and aggressive messages to her post. Julia also notices that many of the harassers are people who don't follow her. She enters the 'Manage Settings' tab, and sets 'Profile Visibility' to 'Followers-only'. Through this, she ensures that her post and the message contained within it can be known to the networked social media audience, while ensuring that malicious users cannot access her profile to harass her, or add harassing responses to her post.

Bystander Scenario

Steven is browsing his social media timeline when he discovers a post from a user called Rachel claiming that "*Lesbians should really stop crushing on straight girls, it's frankly embarrassing how often it happens.*". Many people are criticizing the post to be discriminatory against the LGBTQ+ community, some people even claiming that Rachel is homophobic. However, he notices that the post has an 'Additional Content Added' flag attached to it by Re:SPect. He senses that his perception of the situation might be wrong, and he clicks on the post to see if there are any additional information that could change his perception. He discovers that Rachel has clarified that she identifies as lesbian, and that the post was intended to be a self-deprecating comment on their own situation and not a commentary on how LGBTQ+ people should behave. Steven understands why Rachel made this post, and thinks that the criticism is maybe unjust. He also checks the responses to the post, and notices that many of them are negative, assuming that Rachel is someone who discriminates against sexual minorities. Steven decides to support Rachel by commenting in support of her.

4.4 User Evaluation

To observe the perception of actual Twitter users surrounding our system and suggested features, we conducted a qualitative user study. Due to the potential negative repercussions of measuring for online harassment, we decided against conducting a real-life field study. Instead, we conducted an interview-based study by providing realistic calling out and networked harassment scenarios to our participants and asking them to empathize with the given scenario based on their previous experiences. This type of

speculative study design follows the precedent of anti-harassment systems research such as Squadbox [60] and Unmochon [96].

4.4.1 Methods

The study sessions were conducted through Zoom video call, lasting between 85 and 119 minutes. The session began with a preliminary observation of online harassment-related experiences of the interviewee, and the types of response measures they used to prevent or respond to online harassment. Following this, participants were briefed about the features, design, and usage of Re:SPect, as well as our design motivations. We then moved on to the usage session where participants used Re:SPect to demonstrate their potential responses to online harassment. The usage session was conducted through a think-aloud process, and we conducted a follow-up interview to collect their general feedback about the system as well as insights into the underlying motivations behind the actions they took. Finally, we asked participants about the potential positive and negative effects of implementing anti-harassment features such as in Re:SPect on social media platforms through a semi-structured interview. Participants were paid 30,000 KRW (approx. 23.6 USD)

To observe how participants may use Re:SPect in real harassment scenarios, we provided them with networked harassment/calling out scenarios representing either preemptive or reactive networked harassment situations and asked the participant to demonstrate their response as though they were the victim in the scenario. The preemptive scenario represented a situation where the individual had been called out by the public audience, but at a smaller scale; the reactive scenario presented a situation where the responses were considered more harassing. The scenarios were constructed based on our previous observations of factors that distinguish between online harassment and harmless/valid criticism, such as the scale of calling out and the aggression level of the language used. We note that perceptions of what constitutes harassment may differ between individuals, and thus opted for a more extreme difference in responses between the preemptive and reactive scenarios. The interview results were transcribed for analysis, and then we conducted open coding for themes that emerged from the data.

We interviewed a total of 18 participants, who were recruited through a public Twitter post as well as personal recommendations from previous participants. The condition for participation was Twitter users who have experienced or witnessed networked online harassment in the past 5 years, and those who have speculated about response tactics to mitigate the negative effects of networked online harassment. The participant demographics are detailed in Table 4.2.

4.4.2 Results

In general, participants reacted favorably towards the concept and implementation of Re:SPect that we suggested. Many participants noted that the system would be able to give them “the ability to engage and respond” to harassment instead of being a passive victim, and even encouraging them to speak up more on social media. They were also generally favorable towards the idea of developing anti-harassment tools as they thought the existing measures of responding to harassment were indeed limited. Our results suggest that all three of our initial design goals were met. Here, we organize the common themes that emerged from the interviews and feedback on the system and discuss limitations and potential modes for improvement for Re:SPect.

Table 4.2: Re:SPect User Study Participant Demographics.

ID	Gender	Cisgender/ Transgender	Age	# of Accounts	Harassment Experience	
					Direct Experience	Indirect Experience
P1	F	Cisgender	24	4		O
P2	M	Cisgender	29	2		O
P3	F	Cisgender	24	3		O
P4	F	Cisgender	24	3		O
P5	F	Cisgender	30	3		O
P6	F	Cisgender	24	6	O	
P7	F	Cisgender	23	5	O	
P8	F	Cisgender	28	3		O
P9	F	Cisgender	22	3	O	
P10	F	Cisgender	26	3	O	
P11	F	Cisgender	23	3	O	
P12	Does not wish to answer		24	3		O
P13	F	Cisgender	27	2	O	
P14	Unknown	Transgender	26	2	O	
P15	F	Cisgender	20	1	O	
P16	Does not identify		20	6	O	
P17	F	Cisgender	26	3	O	

A Safety Net from the Negative Effects of Harassment

Many participants thought that they would feel safer with the existence of Re:SPect, especially and even when they are under harassment. The theme of Re:SPect being able to function as a shield (P2, P3, P4, P6, P12, P17) or safety net (P3) from potential harm emerged from several interviews. Several participants (P3, P6, P8, P10, P15) noted that the sheer existence of anti-harassment tactics will make them fear harassment a lot less, causing them a sense of psychological safety.

Even just reducing the stress and fear of harassment helps deal with harassment effectively.

- P15

P10 specifically noted that *“Fear and anxiety comes from the perception that you can’t control the situation - and [this system] gives you exactly that. A sense of control.”*

Allowing for more accurate and efficient information processing Participants mentioned that the summarization feature allowed them to perceive the responses and opinions more clearly. In many cases, participants noted that they would focus on malicious or negative comments even when there were positive or supportive comments, such as in P1’s comment that *“Even if there’s a lot of supportive comments, one bad comment is enough to ruin your mood.”* P9 attributed the fear of harassment to this, saying *“You see one aggressive comment, and then you’re suddenly scared. Because how much of the rest is going to be like that?”* In comparison, the summarization feature organized the responses and opinions by sentiment, and they were able to discover and focus on a lot more of the positive comments, developing a more balanced view of the opinions.

In the context of networked harassment, the ability to bulk process information and provide users with a more condensed view was also viewed favorably in general. Many participants noted that it

is often hard to comprehend the content of the responses in a networked harassment situation due to the scale and content of responses. The traditional method involved reading through each individual comment, mentally processing and compartmentalizing each opinion. This was reported to typically cause fatigue in the participants, leading to them leaning away from trying to understand what is being said. However, as the summary provided a concise view, participants felt less burdened and said that they will check the responses more often as a result (P4, P5, P17).

Protection from immediate exposure to negative responses The response summarization feature (D2) was also mentioned frequently as many participants noted that it provided them with protection from being immediately exposed to negative responses. Many participants noted that on Twitter, there is no way of knowing or expecting what kind of responses you may have before actually checking the responses or QT lists. As these features were also often used for harassment, this led them to feel anxious about checking the responses, as noted in the previous point. However, the summarization feature allowed them to be prepared for the prospect of seeing negative opinions, which participants reported to have reduced the negative psychological impact that the comments had.

I can prepare myself before making the choice to read strongly worded negative comments, thus lessening their impact. - P12

In other cases, participants noted that opinions that were clustered at the lower extreme (that were part of the ‘Overwhelmingly Negative’ category) were often simple vitriol or aggressive opinions that were “*not even worth responding to.*” (P2). As they perceived these opinions and clusters to have minimal communicative or informational value, oftentimes they would simply forego reading this list and not interact with extremely negative opinions at all. Many participants reported that they would gravitate towards looking at the positive comments instead of the negative ones, as they were usually less noticeable. P4 noted that with this feature, they would be able to “*Just think of it as have been controversial, instead of focusing on and remembering the negative comments.*” Several participants (P1, P4, P9, P10, P11) also mentioned that the negative comments were less noticeable in the summary view, allowing them to focus on the positive comments. However, some participants also noted that this type of behavior might cause side effects where people will ignore even valid points of criticism just because they are negative towards them (P2, P3, P5, P6, P8, P9, P14). Even so, some participants, such as P5, P10, and P17, noted that people should speak more carefully and with less aggression to not be categorized in the extreme categories, pointing out that valid critiques are still harassing when the language used is overly aggressive.

Preventing Harassment through Re:SPect

Participants had generally favorable expectations for the potential of Re:SPect in preventing online harassment before it happened. Several participants pointed out that the central problem of online networked harassment is that it causes a sense of helplessness in the user as the extent and scale of the harassment became more than what the individual can handle (P1, P2, P3, P7, P11). P14 appreciated how the features of Re:SPect helped “*match the size of my voice to that of others.*” This was also connected to the importance of timely and appropriate responses as if they missed the ‘golden time’ the harassment would spiral out of control of the individual. However, participants said that with Re:SPect, they will still be able to respond to the harassment after the fact.

Even if I fail in dealing with the harassment in its early stages, [Re:SPect] still gives me a way to fight back. - P3

Protection from exposure to malicious actors Many participants noted that the setting post visibility settings would be useful in preventing harassment from occurring. In the preliminary interview, several participants had said that they would ‘watch their words’ in fear of potentially attracting malicious actors by posting their opinions publicly. Yet, they also recognized their inevitability, comparing malicious actors to ‘traffic accidents’ or ‘natural disasters’, that cannot be avoided no matter how hard one tried. Thus, they speculated that they would use the visibility settings preemptively to prevent any stressful situations, such as P3 who said “*I’ll just put it up every time I feel like I’m saying something remotely divisive, like things I usually use my private accounts for.*” P10 also noted that “*You can reduce the negative psychological impact just by controlling how exposed you are to the public.*”

On the other hand, participants also noted that the visibility settings could help with preventing the spread of harassment after the fact. P2 compared the preemptive and reactive scenarios as stages 2 and 3 respectively (stage 1 referring to a situation with no calling out), and noted that the visibility settings could “*prevent a hypothetical stage 4.*” P7 also mentioned that instead of being helpless to just wait until things simmer down, they will be able to “*actually stop it before it gets out of hand.*”

It gives you a lot more options than just to avoid it altogether, or to just wait. - P7

Participants were also concerned about the possibility of their personal information, usually disclosed in their profiles, being used to abuse them. Several participants (P2, P6, P11, P15) pointed out that they thought that harassment starts or worsens as the conversation moves on from criticizing the action and begins focusing on the flaws of the individual. In this case, participants were enthusiastic about the possibility of the profile visibility setting. P6 noted that they would use this feature to “*Avoid them digging into my previous tweets [or personal life], so that they could condemn me. I want them to focus on the issue at hand.*”

Reducing Possibility of Misunderstandings The post flagging feature was also noted as an important feature that allowed them to actively try to redeem themselves in the face of misunderstandings. Many participants thought that the post flags, especially the ‘Additional Context Added’ flag, were both effective and efficient in terms of responding to the harassment.

I’m just one person, but with this, I can respond to many, many people and express my thoughts to a group of people all at once. - P9

In fact, this feature was almost unanimously praised by the participants as it also reduced the need for them to actively engage with the harassers or callers, which was perceived to be risky as it may instigate further harassment. Participants were enthusiastic about the possibility of “*Preventing my perspectives from being misrepresented or misunderstood.*” (P8) Participants also noted that this feature could prevent harassment from worsening as “*Bystanders would have an easier time catching up on the context*” (P10), allowing people to be less influenced by the ‘flow’ and prevent more people from thoughtlessly participating in harassment without knowing the context.

Harder to Take Responsibility Despite the positive feedback, some participants noted that the features that reduce harassment could actually be used to avoid taking responsibility where they had

actually been in the wrong. This was especially noted in relation to the profile visibility feature. P4 noted that it becomes easier for “*Actual wrongdoers to hide behind anonymity*”, citing examples of sexual violence that were able to be amplified due to the calling out and amplification culture of Twitter. P8 added to this, saying that it is harder to assign responsibility when people can be easily anonymized.

Taking the Initiative to Respond

In general, the existence and features of Re:SPect had the effect of encouraging users to be more active in terms of their responses to harassment. As noted in Table 3.3, responses to harassment could be either active or passive. Participants such as P4, P5, P7, P13, P14, and P16 who had initially said that they would ignore the calling out or delete the Tweet so that they could avoid conflict, said that Re:SPect would allow them to actively respond. P7 also noted that the “*potential range of responses (I can take) is greater*” with Re:SPect. Similarly, P13 highlighted how the system provides a much greater degree of control and a range of possible actions when it comes to dealing with harassment. In general, Re:SPect allowed users to feel safer choosing more active, more engaged responses, while also providing them with increased perceived agency and self-efficacy in the process.

Another element that impacted the perceived agency and self-efficacy was knowing that they made an effort. A sentiment of ‘I know I tried my best’ was repeated across multiple participants (P2, P6, P8), especially after using the post flagging features to denote additional context. Specifically, they felt like a weight was being lifted off their shoulders as they had technically fulfilled the responsibility of clarifying or making an effort to communicate. Thus, once they had already made amends and also made it be known, they were also given more freedom to resent malicious actors - as their explanations would make some attacks clearly over the line.

Once I clarify the misunderstandings, I’ve done my duty. That makes me feel relieved. - P2

Promoting Debate and Discussion in Twitter

Our participants were also optimistic about the potential that Re:SPect could contribute to creating a better space for debate and discussion on Twitter. Several participants, including P9, P13, and P14, were enthusiastic about Re:SPect as they thought the system could contribute to opening up rooms for debate and discussion. They specified the response summary feature as a central factor for this, claiming that being able to see the distribution of opinions will help people form better opinions and also gain a better perspective of others’ opinions. P5, P7, and P16 also noted that the common use of post flags will encourage users to think twice before commenting or QTing others as there may be additional context added later. Finally, P10 noted that allowing users to protect themselves from harassment encourages traditionally discriminated or targeted groups to speak up, enriching the discussion by inviting diverse opinions.

If we protect the users from harassment, then people who were traditionally excluded from these public spaces, minority opinions, can all come together here. Twitter already does that, but that strength could be enhanced even more. - P10

Chapter 5. Discussion

In this section, we discuss the implications of our findings on calling out, online harassment and on social media. Based on these concepts, we also suggest possible design implications to expand upon our current study design, while mitigate the issues surrounding calling out and online harassment.

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 [30]. In this way, Twitter has been used to reverse the power dynamics of media through public sympathy and functioning as a counter-public space [31]. 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 3.3.2). 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 [50]. 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 [97, 50]. 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 [59]. 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 [20, 98]. 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’ [99]. 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 [80, 78].

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 [87]. 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 [100]. Considering the importance and effectiveness of bystander intervention for mitigating online harassment [38, 94, 101, 102, 103, 100], we suggest that platforms should design for possible bystander involvement without fear of such repercussions.

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 [64, 63] and being resigned to the possibility of online hostility [104, 105, 106]. 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 [63].

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 [107]. Sometimes referred to as the ‘Filter Bubble’ [108], 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 [109, 110]. 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 [111, 112, 113], as well as the discussion of how shared behavior is developed within such groups [114], 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 [60, 19, 63]. 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 [115]. We note the necessity of supplementary features such as preemptively blocking new accounts from someone [116] 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 [117], 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 [84], 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 [118]. 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 [119]. 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 [38, 120]. 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 [20] 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 [38, 120, 99]. 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 3.3.1). 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 [21, 121]. 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 [19, 122, 62]. Moreover, as callers and callees differed in their scoping of harassment (Section 3.3.5), 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 [123], 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 [124]. 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 [39]. 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 [76]. 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 Callees' Ability to Engage

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 [81].

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 [81, 19, 20]. 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 [66]. 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 Designing to Prevent Online Harassment

In this thesis, we focused on methods to increase the agency and capacity for response for online harassment victims. However, the needs and perceptions of harassment victims are all unique, and we recognize that there is not a one-size-fits-all solution to online harassment. Thus, there could still be various other methods to design for protecting users against online harassment. Based on our insights from the interview study and design workshop, we expand upon the previous scholarship to suggest additional 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 [66], nor do users agree with these decisions [64, 19]. 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 [20]. 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 paradigms could significantly mitigate the issue surrounding online harassment, and provide scalable and lasting change to improve the social media experience.

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 [70, 125], 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 [77], 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. Features such as locking one’s account or individually blocking harassers’ accounts are used frequently in various social media platforms. 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 [126] 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 [127, 116] and YouTube [128] 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. Mastodon, recently

5.4.3 Providing Indirect Routes for Bystander Intervention

Previous research on bystander intervention focuses on the bystander effect and diffusion of responsibility [129] 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 [129]. 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 [130]. 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, methods such as friendsourced moderation to monitor messages or posts directed to a user could be a useful way to mitigate the direct effects of harassment [60]. 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 [63], 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.

Chapter 6. Conclusion

6.1 Generalizing Across Diverse Social Media Platforms

As of December 2022, the recent discussion surrounding Twitter’s takeover by Tesla CEO Elon Musk [131], as well as his newly introduced management strategies [132] has sparked a significant amount of discussion and change in the social media ecosystem. Some of the major changes include the discourse surrounding the moderation strategies of the platform, its effects on harassment, and of migrating to other platforms, such as Tumblr¹, Mastodon, Hive Social², etc., in place of Twitter. While Twitter’s abundance of online harassment and its moderation policies has been critiqued heavily, the migration sparked more discussion on how this problem persists in other platforms, which again led to a discussion about how moderation and user protection practices would be implemented in newer social media platforms.

For example, Mastodon, a popular alternative post-Twitter due to its similar interface and decentralized management system, was initially applauded due to its limited implementation of ‘amplification’ features. Mastodon does not include a QT-like interface, only containing a ‘Boost’ function that works similarly to that of the RT. The decentralized nature of Mastodon also means that it allows more freedom for users to choose instances with preferable moderation policies, and for server administrators to implement their own standards. However, this has its downfalls - their volunteer workforce is often not enough to provide sufficient response to issues, the administrators’ interests might not always be to protect the harassed user, and the attacks come from across the entire ‘fediverse’ - increasing the challenges in moderation and scale of harassment.

This evidences that many platforms, even with their diverse set of features and use cases, often face similar problems in moderation and networked harassment. While the current study focused on the experiences of Twitter users and their experiences, we believe that the implications noted in this thesis can be generalizable to other platforms as well.

6.2 Limitations and Future Work

In our interview study, 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 for both the interview study and design workshop. 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

¹<https://www.tumblr.com/>

²<https://www.hivesocial.app/>

have influenced or limited the reach of the Tweet. 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.

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.

Bibliography

- [1] Sarita Yardi and danah boyd. Dynamic debates: An analysis of group polarization over time on twitter. *Bulletin of science, technology & society*, 30(5):316–327, 2010.
- [2] Julian Ausserhofer and Axel Maireder. National politics on twitter: Structures and topics of a networked public sphere. *Information, communication & society*, 16(3):291–314, 2013.
- [3] Shuzhe Yang, Anabel Quan-Haase, and Kai Rannenberg. The changing public sphere on twitter: Network structure, elites and topics of the #righttobeforgotten. *New media & society*, 19(12):1983–2002, 2017.
- [4] Amy X. Zhang and Scott Counts. *Modeling Ideology and Predicting Policy Change with Social Media: Case of Same-Sex Marriage*, page 2603–2612. Association for Computing Machinery, New York, NY, USA, 2015.
- [5] Zhe Liu and Ingmar Weber. Is twitter a public sphere for online conflicts? a cross-ideological and cross-hierarchical look. In *International Conference on Social Informatics*, pages 336–347. Springer, 2014.
- [6] Elanor Colleoni, Alessandro Rozza, and Adam Arvidsson. Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. *Journal of communication*, 64(2):317–332, 2014.
- [7] Carly Gieseler. *The Voices of #MeToo: From Grassroots Activism to a Viral Roar*. Rowman & Littlefield, July 2019.
- [8] Munmun De Choudhury, Shagun Jhaver, Benjamin Sugar, and Ingmar Weber. Social media participation in an activist movement for racial equality. In *Tenth International AAAI Conference on Web and Social Media*, pages 92–101, 2016.
- [9] Rajesh Basak, Niloy Ganguly, Shamik Sural, and Soumya K. Ghosh. 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*, pages 11–12, Montréal, Québec, Canada, 2016. ACM Press.
- [10] Jon Ronson. How One Stupid Tweet Blew Up Justine Sacco’s Life. *The New York Times*, February 2015.
- [11] Rajesh Basak, Shamik Sural, Niloy Ganguly, and Soumya K. Ghosh. Online Public Shaming on Twitter: Detection, Analysis, and Mitigation. *IEEE Transactions on Computational Social Systems*, 6(2):208–220, April 2019.
- [12] Joseph Ching Velasco. You are Cancelled: Virtual Collective Consciousness and the Emergence of Cancel Culture as Ideological Purging. *rupkatha*, 12(5), October 2020.
- [13] Antara Kashyap. Cancel Culture: Threat to Freedom of Expression or a Form of Accountability?, April 2021.

- [14] Eve Ng. No Grand Pronouncements Here...: Reflections on Cancel Culture and Digital Media Participation. *Television & New Media*, 21(6):621–627, September 2020.
- [15] Kaitlynn Mendes, Jessica Ringrose, and Jessalynn Keller. #MeToo and the promise and pitfalls of challenging rape culture through digital feminist activism. *European Journal of Women's Studies*, 25(2):236–246, May 2018.
- [16] Lisa Nakamura. The unwanted labour of social media: Women of colour call out culture as venture community management. *New Formations*, 86(86):106–112, 2015.
- [17] Gwen Bouvier. Racist call-outs and cancel culture on Twitter: The limitations of the platform's ability to define issues of social justice. *Discourse, Context & Media*, 38:100431, December 2020.
- [18] Jon Ronson. *So You've Been Publicly Shamed*. Riverhead Books, December 2016.
- [19] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. Online Harassment and Content Moderation: The Case of Blocklists. *ACM Trans. Comput.-Hum. Interact.*, 25(2):1–33, April 2018.
- [20] Alice E Marwick. Morally motivated networked harassment as normative reinforcement. *Social Media+ Society*, 7(2):20563051211021378, 2021.
- [21] Colette Langos. Cyberbullying: The Challenge to Define. *Cyberpsychology, Behavior, and Social Networking*, 15(6):285–289, June 2012.
- [22] Meredith D. Clark. DRAG THEM: A brief etymology of so-called “cancel culture”. *Communication and the Public*, 5(3-4):88–92, September 2020.
- [23] André Brock Jr. *Distributed Blackness: African American Cybercultures*. NYU Press, February 2020.
- [24] Pippa Norris. Cancel culture: Myth or reality? *Political Studies*, page 00323217211037023, 2021.
- [25] Gwen Bouvier and David Machin. What gets lost in Twitter ‘cancel culture’ hashtags? Calling out racists reveals some limitations of social justice campaigns. *Discourse & Society*, 32(3):307–327, May 2021.
- [26] Brian L Ott. The age of twitter: Donald j. trump and the politics of debasement. *Critical studies in media communication*, 34(1):59–68, 2017.
- [27] Lindsay Ellis. Mask off, April 2021.
- [28] Michał Krzyżanowski and Per Ledin. Uncivility on the web: Populism in/and the borderline discourses of exclusion. *Journal of Language and Politics*, 16(4):566–581, 2017.
- [29] David Theo Goldberg. *Are we all postracial yet?* John Wiley & Sons, 2015.
- [30] Yarimar Bonilla and Jonathan Rosa. #ferguson: Digital protest, hashtag ethnography, and the racial politics of social media in the united states. *American ethnologist*, 42(1):4–17, 2015.
- [31] Michael Salter. Justice and revenge in online counter-publics: Emerging responses to sexual violence in the age of social media. *Crime, Media, Culture*, 9(3):225–242, 2013.

- [32] Robert Asen. Seeking the “counter” in counterpublics. *Communication theory*, 10(4):424–446, 2000.
- [33] Bianca Fileborn. Justice 2.0: Street harassment victims’ use of social media and online activism as sites of informal justice. *British journal of criminology*, 57(6):1482–1501, 2017.
- [34] Sanja Milivojevic and Alyce McGovern. The death of jill meagher: Crime and punishment on social media. *International journal for crime, justice and social democracy*, 3(3):22–39, 2014.
- [35] Amit M Schejter and Noam Tirosh. “seek the meek, seek the just”: Social media and social justice. *Telecommunications policy*, 39(9):796–803, 2015.
- [36] Inbal Peleg-Koriat and Carmit Klar-Chalamish. The #MeToo movement and restorative justice: exploring the views of the public. *Contemporary Justice Review*, 23(3):239–260, July 2020.
- [37] Sarita Schoenebeck, Carol F. Scott, Emma Grace Hurley, Tammy Chang, and Ellen Selkie. Youth Trust in Social Media Companies and Expectations of Justice: Accountability and Repair After Online Harassment. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1):1–18, April 2021.
- [38] Lindsay Blackwell, Tianying Chen, Sarita Schoenebeck, and Cliff Lampe. When Online Harassment Is Perceived as Justified. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1):10, 2018.
- [39] Sarita Schoenebeck, Oliver L Haimson, and Lisa Nakamura. Drawing from justice theories to support targets of online harassment. *New Media & Society*, 23(5):1278–1300, May 2021.
- [40] Hanlin Li, Disha Bora, Sagar Salvi, and Erin Brady. Slacktivists or Activists?: Identity Work in the Virtual Disability March. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13, Montreal QC Canada, April 2018. ACM.
- [41] Maya Stewart and Ulrike Schultze. Producing solidarity in social media activism: The case of My Stealthy Freedom. *Information and Organization*, 29(3):100251, September 2019.
- [42] Anthony McCosker. Social Media Activism at the Margins: Managing Visibility, Voice and Vitality Affects. *Social Media + Society*, 1(2):205630511560586, July 2015.
- [43] Yu-Hao Lee and Gary Hsieh. 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*, pages 811–820, 2013.
- [44] Ariadne Vromen. Digital citizenship and political engagement. In *Digital citizenship and political engagement*, pages 9–49. Springer, 2017.
- [45] Max Halupka. Clicktivism: A Systematic Heuristic. *Policy & Internet*, 6(2):115–132, 2014.
- [46] Max Halupka. The legitimisation of clicktivism. *Australian Journal of Political Science*, 53(1):130–141, January 2018.
- [47] Mathias Klang and Nora Madison. Vigilantism or outrage: An exploration of policing social norms through social media. *Ethics for a Digital Age*, 2:151–165, 2018.
- [48] Michael Wenzel and Tyler G Okimoto. Retributive justice. In *Handbook of social justice theory and research*, pages 237–256. Springer, 2016.

- [49] Kate Klonick. A new taxonomy for online harms. *Boston University Law Review Annex*, 95:53–55, 2015.
- [50] danah boyd. 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):119–142, 2008.
- [51] Robert Slonje and Peter K Smith. Cyberbullying: Another main type of bullying? *Scandinavian journal of psychology*, 49(2):147–154, 2008.
- [52] Alisdair A Gillespie. Cyber-bullying and harassment of teenagers: The legal response. *Journal of Social Welfare & Family Law*, 28(2):123–136, 2006.
- [53] Paul Benjamin Lowry, Jun Zhang, Chuang Wang, and Mikko Siponen. 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):962–986, 2016.
- [54] John Suler. The online disinhibition effect. *Cyberpsychology & behavior*, 7(3):321–326, 2004.
- [55] Robyn M Cooper and Warren J Blumenfeld. Responses to cyberbullying: A descriptive analysis of the frequency of and impact on lgbt and allied youth. *Journal of LGBT Youth*, 9(2):153–177, 2012.
- [56] Francine Dehue, Catherine Bolman, and Trijntje Völlink. Cyberbullying: Youngsters’ experiences and parental perception. *CyberPsychology & Behavior*, 11(2):217–223, 2008.
- [57] Collin Gifford Brooke. Forgetting to be (post) human: Media and memory in a kairotic age. *JAC*, pages 775–795, 2000.
- [58] Viktor Mayer-Schönberger. *Delete: The virtue of forgetting in the digital age*. Princeton University Press, 2011.
- [59] Michael S Bernstein, Eytan Bakshy, Moira Burke, and Brian Karrer. Quantifying the invisible audience in social networks. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 21–30, New York, NY, USA, 2013. Association for Computing Machinery.
- [60] Kaitlin Mahar, Amy X. Zhang, and David Karger. Squadbox: A Tool to Combat Email Harassment Using Friendsourced Moderation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13, Montreal QC Canada, April 2018. ACM.
- [61] Ersilia Menesini and Annalaura Nocentini. Cyberbullying Definition and Measurement: Some Critical Considerations. *Zeitschrift für Psychologie / Journal of Psychology*, 217(4):230–232, January 2009.
- [62] Jane Im, Jill Dimond, Melody Berton, Una Lee, Katherine Mustelier, Mark S. Ackerman, and Eric Gilbert. 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*, pages 1–18, Yokohama Japan, May 2021. ACM.

- [63] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW):1–19, December 2017.
- [64] Jessica Vitak, Kalyani Chadha, Linda Steiner, and Zahra Ashktorab. 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*, pages 1231–1245, Portland Oregon USA, February 2017. ACM.
- [65] Ysabel Gerrard. Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society*, 20(12):4492–4511, 2018.
- [66] Jessica A Pater, Moon K Kim, Elizabeth D Mynatt, and Casey Fiesler. Characterizations of online harassment: Comparing policies across social media platforms. In *Proceedings of the 19th international conference on supporting group work*, pages 369–374, 2016.
- [67] Sarah Myers West. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, 20(11):4366–4383, 2018.
- [68] Ganaele Langlois, Greg Elmer, Fenwick McKelvey, and Zachary Devereaux. Networked publics: The double articulation of code and politics on facebook. *Canadian Journal of Communication*, 34(3), 2009.
- [69] J Nathan Matias, Amy Johnson, Whitney Erin Boesel, Brian Keegan, Jaclyn Friedman, and Charlie DeTar. Reporting, reviewing, and responding to harassment on twitter. *arXiv preprint arXiv:1505.03359*, 2015.
- [70] George Kennedy, Andrew McCollough, Edward Dixon, Alexei Bastidas, John Ryan, Chris Loo, and Saurav Sahay. Technology solutions to combat online harassment. In *Proceedings of the first workshop on abusive language online*, pages 73–77, 2017.
- [71] Jennifer Golbeck, Zahra Ashktorab, Rashad O Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A Geller, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, et al. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on web science conference*, pages 229–233, 2017.
- [72] Joshua Guberman, Carol Schmitz, and Libby Hemphill. Quantifying toxicity and verbal violence on twitter. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*, pages 277–280, 2016.
- [73] Thabo Mahlangu, Chunling Tu, and Pius Owolawi. A review of automated detection methods for cyberbullying. In *2018 International Conference on Intelligent and Innovative Computing Applications (ICONIC)*, pages 1–5. IEEE, 2018.
- [74] Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. Tackling online abuse: A survey of automated abuse detection methods. *arXiv preprint arXiv:1908.06024*, 2019.
- [75] Austin P Wright, Omar Shaikh, Haekyu Park, Will Epperson, Muhammed Ahmed, Stephane Pinel, Duen Horng Chau, and Diyi Yang. Recast: Enabling user recourse and interpretability of toxicity detection models with interactive visualization. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–26, 2021.

- [76] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 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, CSCW '17*, page 1217–1230, New York, NY, USA, 2017. Association for Computing Machinery.
- [77] Travis Kriplean, Jonathan Morgan, Deen Freelon, Alan Borning, and Lance Bennett. Supporting reflective public thought with considerit. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 265–274, New York, NY, USA, 2012. Association for Computing Machinery.
- [78] Travis Kriplean, Michael Toomim, Jonathan Morgan, Alan Borning, and Amy Ko. Is this what you meant? promoting listening on the web with reflect. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1559–1568, New York, NY, USA, 2012. Association for Computing Machinery.
- [79] Matti Nelimarkka, Jean Philippe Rancy, Jennifer Grygiel, and Bryan Semaan. (re) design to mitigate political polarization: Reflecting habermas’ ideal communication space in the united states of america and finland. *Proceedings of the ACM on Human-computer Interaction*, 3(CSCW):1–25, 2019.
- [80] Hyunwoo Kim, Haesoo Kim, Kyung Je Jo, and Juho Kim. Starrythoughts: Facilitating diverse opinion exploration on social issues. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–29, 2021.
- [81] Rebecca Lewis, Alice E Marwick, and William Clyde Partin. “we dissect stupidity and respond to it”: Response videos and networked harassment on youtube. *American Behavioral Scientist*, 65(5):735–756, 2021.
- [82] Yu Wang, Jiebo Luo, Richard Niemi, Yuncheng Li, and Tianran Hu. Catching fire via ”likes”: Inferring topic preferences of trump followers on twitter. In *Tenth International AAAI Conference on Web and Social Media*, pages 719–722, 2016.
- [83] danah boyd, Scott Golder, and Gilad Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *2010 43rd Hawaii international conference on system sciences*, pages 1–10. IEEE, 2010.
- [84] Kiran Garimella, Ingmar Weber, and Munmun De Choudhury. Quote rts on twitter: usage of the new feature for political discourse. In *Proceedings of the 8th ACM Conference on Web Science*, pages 200–204, New York, NY, USA, 2016. Association for Computing Machinery.
- [85] Hemant Purohit, Andrew Hampton, Valerie L Shalin, Amit P Sheth, John Flach, and Shreyansh Bhatt. What kind of #conversation is twitter? mining #psycholinguistic cues for emergency coordination. *Computers in Human Behavior*, 29(6):2438–2447, 2013.
- [86] Urban Dictionary: Lrt.
- [87] Fernando Domínguez-Hernández, Lars Bonell, and Alejandro Martínez-González. A systematic literature review of factors that moderate bystanders’ actions in cyberbullying. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 12(4), 2018.

- [88] Rob Kling, Ya-ching Lee, Al Teich, and Mark S Frankel. Assessing anonymous communication on the internet: Policy deliberations. *The Information Society*, 15(2):79–90, 1999.
- [89] Urban Dictionary: ratioed.
- [90] Kiel Long, John Vines, Selina Sutton, Phillip Brooker, Tom Feltwell, Ben Kirman, Julie Barnett, and Shaun Lawson. ” could you define that in bot terms”? requesting, creating and using bots on reddit. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3488–3500, 2017.
- [91] Joseph Seering, Tianmi Fang, Luca Damasco, Mianhong’Cherie’ Chen, Likang Sun, and Geoff Kaufman. Designing user interface elements to improve the quality and civility of discourse in online commenting behaviors. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019.
- [92] Aparajita Bhandari, Marie Ozanne, Natalya N Bazarova, and Dominic DiFranzo. Do you care who flagged this post? effects of moderator visibility on bystander behavior. *Journal of Computer-Mediated Communication*, 26(5):284–300, 2021.
- [93] Samuel Hardman Taylor, Dominic DiFranzo, Yoon Hyung Choi, Shruti Sannon, and Natalya N Bazarova. Accountability and empathy by design: Encouraging bystander intervention to cyberbullying on social media. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–26, 2019.
- [94] Dominic DiFranzo, Samuel Hardman Taylor, Francesca Kazerooni, Olivia D Wherry, and Natalya N Bazarova. Upstanding by design: Bystander intervention in cyberbullying. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–12, New York, NY, USA, 2018. Association for Computing Machinery.
- [95] Una Lee and Dann Toliver. Building consentful tech. 2017, 2017.
- [96] Sharifa Sultana, Mitrasree Deb, Ananya Bhattacharjee, Shaid Hasan, SM Raihanul Alam, Trishna Chakraborty, Prianka Roy, Samira Fairuz Ahmed, Aparna Moitra, M Ashraf Amin, et al. ‘unmochon’: A tool to combat online sexual harassment over facebook messenger. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2021.
- [97] Alice E Marwick and danah boyd. I tweet honestly, i tweet passionately: Twitter users, context collapse, and the imagined audience. *New media & society*, 13(1):114–133, 2011.
- [98] William J Brady, Julian A Wills, John T Jost, Joshua A Tucker, and Jay J Van Bavel. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28):7313–7318, 2017.
- [99] Ann DeSmet, Sara Bastiaensens, Katrien Van Cleemput, Karolien Poels, Heidi Vandebosch, Greet Cardon, and Ilse De Bourdeaudhuij. 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*, 57:398–415, 2016.
- [100] Lisa M Jones, Kimberly J Mitchell, and Heather A Turner. Victim reports of bystander reactions to in-person and online peer harassment: A national survey of adolescents. *Journal of youth and adolescence*, 44(12):2308–2320, 2015.

- [101] Miia Sainio, René Veenstra, Gijs Huitsing, and Christina Salmivalli. Victims and their defenders: A dyadic approach. *International journal of behavioral development*, 35(2):144–151, 2011.
- [102] Nicholas Brody. Bystander intervention in cyberbullying and online harassment: The role of expectancy violations. *International Journal of Communication*, 15:21, 2021.
- [103] Nicholas Brody and Anita L Vangelisti. Bystander intervention in cyberbullying. *Communication Monographs*, 83(1):94–119, 2016.
- [104] Martin EP Seligman. Learned helplessness. *Annual review of medicine*, 23(1):407–412, 1972.
- [105] Amanda Burgess-Proctor, Justin W Patchin, and Sameer Hinduja. Cyberbullying and online harassment: Reconceptualizing the victimization of adolescent girls. *Female crime victims: Reality reconsidered*, pages 153–175, 2009.
- [106] Kususanto Prihadi, Yen Ling Hui, Melissa Chua, and Calvin KW Chang. Cyber-victimization and perceived depression: Serial mediation of self-esteem and learned-helplessness. *International Journal of Evaluation and Research in Education*, 8(4):563–574, 2019.
- [107] Dominic Spohr. Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *Business Information Review*, 34(3):150–160, 2017.
- [108] E. Pariser. *The Filter Bubble: What The Internet Is Hiding From You*. Penguin Books Limited, 2011.
- [109] Elisa Jayne Bienenstock, Phillip Bonacich, and Melvin Oliver. The effect of network density and homogeneity on attitude polarization. *Social Networks*, 12(2):153–172, 1990.
- [110] Michael A Hogg, John C Turner, and Barbara Davidson. Polarized norms and social frames of reference: A test of the self-categorization theory of group polarization. *Basic and Applied Social Psychology*, 11(1):77–100, 1990.
- [111] Martin Grandjean. A social network analysis of twitter: Mapping the digital humanities community. *Cogent Arts & Humanities*, 3(1):1171458, 2016.
- [112] Kwan Hui Lim and Amitava Datta. Finding twitter communities with common interests using following links of celebrities. In *Proceedings of the 3rd international workshop on Modeling social media*, pages 25–32, 2012.
- [113] Abraham Ronel Martínez Teutle. Twitter: Network properties analysis. In *2010 20th International Conference on Electronics Communications and Computers (CONIELECOMP)*, pages 180–186, 2010.
- [114] Logan Molyneux and Rachel R Mourão. Political journalists’ normalization of twitter: Interaction and new affordances. *Journalism Studies*, 20(2):248–266, 2019.
- [115] Fayika Farhat Nova, Michael Ann DeVito, Pratyasha Saha, Kazi Shohanur Rashid, Shashwata Roy Turzo, Sadia Afrin, and Shion Guha. ” facebook promotes more harassment” social media ecosystem, skill and marginalized hijra identity in bangladesh. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–35, 2021.
- [116] Introducing new tools to protect our community from abuse | Instagram Blog, 4 2021.

- [117] Eun-mee Kim and Jennifer Ihm. More than virality: Online sharing of controversial news with activated audience. *Journalism & Mass Communication Quarterly*, 97(1):118–140, 2020.
- [118] Twitter Support. 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>, August 2020.
- [119] Jiyoun Kim. The meaning of numbers: Effect of social media engagement metrics in risk communication. *Communication Studies*, 72(2):195–213, 2021.
- [120] Ann DeSmet, Sara Bastiaensens, Katrien Van Cleemput, Karolien Poels, Heidi Vandebosch, and Ilse De Bourdeaudhuij. Mobilizing bystanders of cyberbullying: an exploratory study into behavioural determinants of defending the victim. *Annual review of cybertherapy and telemedicine*, 10:58–63, 2012.
- [121] Peter K Smith, Cristina Del Barrio, and Robert S Tokunaga. Definitions of bullying and cyberbullying: How useful are the terms? In *Principles of cyberbullying research*, pages 54–68. Routledge, 2012.
- [122] Shagun Jhaver, Larry Chan, and Amy Bruckman. The view from the other side: The border between controversial speech and harassment on kotaku in action. *First Monday*, 23(2), 2018.
- [123] Casey Fiesler, Cliff Lampe, and Amy S Bruckman. 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*, pages 1450–1461, New York, NY, USA, 2016. Association for Computing Machinery.
- [124] Alice Marwick and danah boyd. ‘it’s just drama’: Teen perspectives on conflict and aggression in a networked era. *Journal of youth studies*, 17(9):1187–1204, 2014.
- [125] Tolba Marwa, Ouadfel Salima, and Meshoul Souham. Deep learning for online harassment detection in tweets. In *2018 3rd International Conference on Pattern Analysis and Intelligent Systems (PAIS)*, pages 1–5. IEEE, 2018.
- [126] Twitter Safety. Your Tweets = Your space. Now you can change who can reply to you even after you Tweet. <https://t.co/3HFSjAotg7>, July 2021.
- [127] Giving People More Control on Instagram and Facebook, 5 2021.
- [128] Change your subscription privacy settings - YouTube Help.
- [129] Randy Yee Man Wong, Christy MK Cheung, Bo Xiao, and Jason Bennett Thatcher. Standing up or standing by: Understanding bystanders’ proactive reporting responses to social media harassment. *Information Systems Research*, 32(2):561–581, 2021.
- [130] Kate Crawford and Tarleton Gillespie. What is a flag for? social media reporting tools and the vocabulary of complaint. *New Media & Society*, 18(3):410–428, 2016.
- [131] Kate Cogner and Lauren Hirsch. Elon Musk Completes \$44 Billion Deal to Own Twitter - The New York Times. *The New York Times*, October 2022.

[132] Faiz Siddiqui and Jeremy B. Merrill. Musk issues ultimatum to staff: Commit to ‘hardcore’ Twitter or take severance. *Washington Post*, November 2022.

Acknowledgment

First of all, I thank my advisors: Prof. Juho Kim and Prof. Jeong-woo Jang, for the amazing advice, support and opportunities they have provided me during my time here. I have been truly blessed to receive your teaching, and I learned so much in the process.

I thank my friends and collaborators at KIXLAB: Juhoon, Hyunwoo, Jeongeon, Yoonseo, and countless others who provided me with the most amazing and heartwarming 2+ years. I do not exaggerate when I say I would not have been able to do this without you. I will miss our shared all-nighters, coffee chats, the inside jokes and even the work... but most of all, I will miss you guys.

I thank my family for the support and love they have sent me, and for giving me the privilege and ability to be able to pursue what I love.

I thank Soomin for always being there for me, and believing in me even in times when I couldn't - and always giving me the joy and strength to persevere.

I thank HaeEun, my sister, for being an amazing collaborator and the biggest inspiration throughout my life and research career. You are (literally) the reason that I exist. I love you, and as always, I wish you all the best from the bottom of my heart.

Finally, I thank everyone that helped me and supported me throughout this journey. My study participants, interviewees, everyone who has ever retweeted my call for participants, past collaborators, online and offline friends who have listened to many a tirade and yet still supported me... I will forever be grateful.

P.S. Special thanks for SEVENTEEN for the moral support and strength throughout the two years. Here's to hoping that I don't regret adding this note in the future. :)

Curriculum Vitae in Korean

이 름: 김 해 수
생 년 월 일: 1997년 11월 08일
전 자 주 소: haesookim@kaist.ac.kr

학 력

2013. 3. – 2016. 2. 고양외국어고등학교
2016. 2. – 2021. 2. 서울대학교 자유전공학부 (학사)
2021. 3. – 2023. 2. 한국과학기술원 전산학부 (석사)

경 력

2020. 3. – 2020. 6. 서울대학교 기초교육원 조교
2020. 9. – 2021. 1. (주) 그라인더 제품 디자이너
2021. 3. – 2021. 12. 한국과학기술원 전산학부 조교

연구 업 적

1. **Haesoo Kim**, Haeun Kim, Juho Kim, and Jeong-woo Jang, “When Does it Become Harassment?: An Investigation of Online Criticism and Calling Out in Twitter,” *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2) (CSCW ’22), ACM
2. Taewan Kim, **Haesoo Kim**, Hayeon Lee, Hwarang Goh, Shakhboz Abdigapporov, Mingon Jeong, Hyunsung Cho, Kyungsik Han, Youngtae Noh, Sung-Ju Lee, and Hwajung Hong, “Prediction for Retrospection: Integrating Algorithmic Stress Prediction into Personal Informatics Systems for College Students’ Mental Health,” *In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI ’22)*, ACM
3. Mina Huh, Yunjung Lee, Dasom Choi, **Haesoo Kim**, Uran Oh, and Juho Kim, “Cocomix: Leveraging Comments to Improve Webtoon Accessibility for Blind or Low Vision Readers,” *In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI ’22)*, ACM
4. Hyunwoo Kim, **Haesoo Kim**, Kyung Je Jo, and Juho Kim, “StarryThoughts: Facilitating Diverse Opinion Exploration on Social Issues,” *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1) (CSCW ’22), ACM