Re:SPect: Enabling Active and Scalable Responses to Networked Online Harassment

HAESOO KIM, School of Computing, KAIST, Republic of Korea
JUHOON LEE, School of Computing, KAIST, Republic of Korea
JUHO KIM, School of Computing, KAIST, Republic of Korea
JEONG-WOO JANG, School of Digital Humanities and Computational Social Sciences, KAIST, Republic of Korea

Online harassment, especially networked harassment at scale, has become an increasingly serious issue that pervades many social media platforms. In this study, we investigated the nature and harms of networked harassment on Twitter through design workshops (n = 11) and developed a set of design goals focusing on empowering the individual to fight back against harassment. We designed Re:SPect, an anti-harassment tool promoting scalable and active responses to networked harassment. We evaluated Re:SPect through a simulated scenario-based study with Twitter users (n = 18) who had directly or indirectly experienced networked harassment. Our findings reveal that users felt safer and more empowered as Re:SPect enabled them to manage interactions with a larger audience. Users felt less anxious about the potential of being harassed, while the summarization features of Re:SPect allowed users to perceive the situation more objectively. Based on the findings, we discuss how Re:SPect's features could be utilized in promoting healthier online discussion, as well as theoretical implications in designing such systems.

CCS Concepts: • **Human-centered computing → Social networking sites**.

Additional Key Words and Phrases: online harassment, networked harassment, calling out, social media, Twitter, anti-harassment systems

ACM Reference Format:
Haesoo Kim, Juhoon Lee, Juho Kim, and Jeong-woo Jang. 2018. Re:SPect: Enabling Active and Scalable Responses to Networked Online Harassment. In . ACM, New York, NY, USA, 30 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Online harassment is a rising problem in the social media ecosystem. As social media use takes up an increasingly significant part of our daily lives and social interactions, recognizing and moderating harassment has become a core task for social media platforms. Users often undergo extreme psychological duress when faced with online harassment, sometimes even changing their online behavior [63], or migrating away from social media platforms [33, 74]. This often results in a decline in the quality of discourse on these platforms [17, 79, 93]. Even when they have not experienced online harassment personally, many users are anxious about potentially attracting harassment, leading them to preemptively change their behaviors [60].

A particular type of harassment that has been gaining prominence is 'networked harassment' [66]. Networked harassment refers to a case where harassment, often at scale, is encouraged or instigated by members of an interconnected online network. This is different from the traditional dyadic notion

CSCW'23, 2023, Woodstock, NY
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06...$15.00
https://doi.org/XXXXXXX.XXXXXXX

of bullying or harassment where one harasser repeatedly engages with the victim in an unwanted fashion [61, 75]. In networked harassment, each individual's contributions may be minimal, but the experience of harassment could feel much worse as they are subject to more overall negative comments. This distributive nature means networked harassment does not map easily to traditional or legal models of harassment, and thus becomes much harder to intervene [56].

Most social media systems utilize user reports [68, 104] and blocking to allow users to respond to offending users [19, 88]. However, users' perception of harassment may be different from what the policies dictate [37, 75], reducing the perceived effectiveness of reporting. These are also tedious as each user has to block or report each individual offender – leading to community efforts, such as blocklists, to apply these solutions at scale [35, 46]. A significant body of work in HCI and CSCW has explored methods to detect [40], prevent [58, 92, 105] and resolve harassment [6, 24, 98]. However, many studies focus more on specifying and stopping the offender instead of explicitly considering the needs of victims. Systems such as Unmochon [96] and Squadbox [64] have aimed to provide practical response measures for victims of harassment through shame-based justice or receiving the help of friends in email moderation. Even so, there is still a relative lack of scholarship on enabling victims to respond and take action against networked harassment at scale.

In this paper, we approach this by providing victims of networked harassment with the ability to manage responses at scale. Specifically, we focus on networked harassment on Twitter, as the platform's emphasis on amplification and abundance of harassment [35, 46] makes it a representative space to study its effects. Through an iterative design process, we evaluated our initial system concept through three design workshop sessions with Twitter users (n = 11), aiming to understand the key factors and considerations when developing systems to prevent networked harassment. The findings revealed that users want to efficiently communicate the intent of the original poster (OP) while reducing stress by distancing themselves from the events and preventing the amplification of the post.

Building from these insights, we present Re:SPect, a system that provides victims of harassment with ways to protect themselves from networked harassment and amplification in open online communication. Re:SPect includes features such as (i) granular visibility controls, (ii) cluster-based summarization of responses to the post, and (iii) mass response measures (e.g. post flags) that emphasize additional context posted by the OP. To evaluate the potential effects of Re:SPect, we conducted a simulated scenario-based user study with 18 Twitter users who had either directly or indirectly experienced networked harassment on Twitter. Our results showed that participants valued Re:SPect's features that allowed them to distance themselves from the harassment, both practically and emotionally. Participants also reported that with Re:SPect, they would more actively protect their reputation and self rather than passively wait for the situation to resolve itself. Based on the findings, we discuss the implications for designing social media platforms to reduce harassment, promoting psychological safety and agency in their users, and how this might be utilized in platforms other than Twitter.

2 RELATED WORK

2.1 Online and Networked Harassment in Social Media

Online harassment refers to a broad range of deviant behaviors that target an individual, including but not limited to hate speech, doxxing, cyberstalking, and physical threats [26, 75, 89]. Previous examinations of online harassment often refer to qualities such as malicious intent, aggressive behavior, and negative impact towards a specific recipient [9, 26, 56, 60], borrowing upon more traditional definitions of bullying and harassment. However, due to the ever-evolving dynamics of the online space, perspectives on what constitutes harassment are prone to change, and may also differ according to the user or group [46]. Thus, existing definitions of harassment are often insufficient in summarizing the diverse experiences of social media users, introducing challenges in policy-making and moderation [56].

An example of one such challenge is the issue of scale. The open nature of online spaces often leads to conversations of a much bigger scale than in real life, which has the potential to attract aggressors or unwanted incidents. Jhaver et al. have previously observed how scale-based tactics such as brigading or dogpiling are utilized by harassers or antagonize their victims [46].

More recent literature defines this type of harassment at scale as 'networked harassment'. Marwick et al. describe networked harassment as a situation where a networked group of individuals organizes against an individual to antagonize them [61, 67]. This increased scale often exacerbates harassment by amplifying elements such as power differentials [70]. persistence [12, 48] , and visibility [5, 101]. The anonymity in numbers also encourages deviant behavior, as the reduced sense of accountability causes disinhibition and deindividuation [62], coupled with the fact that the networked audiences are often left invisible [12]

In such cases, harassment can stem even from benign intent, or with the consensus that it is justified [8] or morally correct [66]. Individuals also may not understand the implications of their actions when considered at scale [52]. This goes hand-in-hand with 'canceling' or 'calling out' behavior, in which an audience publicly shames an individual for perceived wrongful action [23, 47, 52, 72]. While the individual critiques could come from a genuine place of concern or justification, it still has the potential to harm its recipient as it is accumulated with a multitude of other comments [52, 61, 66, 83]. This sense of morality further blurs the line between harassment and valid criticism, especially at the individual scale. This makes it harder for perpetrators to understand the consequences of their actions [52], and even harder for platforms to address these issues.

We build upon this perspective that classifying behaviors as online harassment doesn't necessarily require malicious intent, but could also be defined by scale. That is, harassment can happen if the scale of responses is more than what one could manage [43, 46, 52]. Thus, it is important to also explore definitions and classifications of harassment not only in 'what' the perpetrator did but rather in how the target perceived the action and the impact it had on them [9, 21, 22, 60].In this paper, we focus on this broader perspective of online harassment, exploring methods to protect users from both intentional and unintentional harms and making a more inclusive online space.

2.1.1 Networked Harassment in Twitter. The paper specifically focuses on networked harassment on Twitter. Twitter is a platform notorious for its prevalence of online harassment [46] and calling out behaviors [11, 23], serving as an ideal ground for networked harassment thanks to its amplification features [10]. Recently, the Quote Tweet function, a feature that allows users to repost a tweet while adding their own comments [34], is considered as significantly correlated to the spread of harassment on Twitter [52]. This is because the added exposure opens up room for even more amplification [80], and for outgroups to share their negative opinions [95, 103]. We focus especially on the types of networked harassment often instigated or exacerbated through the use of such features, while recognizing that these are not the only pathways through which harassment occurs.

It is important to note that neither the practice of public critique nor the Quote Tweet feature are inherently harmful. Scholars have previously noted the importance of open communication features, especially how they introduce criticism and increase the visibility of marginalized identities [85–87]. Flowers has previously noted the specific role of features such as Quote Tweets in the formation of counterpublics [32]. However, the public visibility and increased engagement also attract more negative attention, often making it a 'numbers game' and potentially exacerbating harassment [52, 85, 95]. The counterpublic strategies may even be appropriated by mainstream networks, as in the case of cancel culture, to further oppress marginalized groups [1]. Considering this, our work aims to provide alternative methods of communication that maintains the benefit of open communication while preventing the potential negative repercussions that occur in existing platforms.

2.2 Combating Online Harassment

In response to the pervasiveness of abusive behavior, most social media platforms adopt some form of content moderation to protect their users through automatic detection, user reports, and more [36, 82]. Platform interventions include suspending the offender's account or deleting the post, or more personal solutions such as blocking or muting that individual [46, 75]. However, despite these efforts, users often do not trust social media platforms' ability to achieve a fair resolution to harassment [90].

NLP research has explored methods to automatically detect harassment, such as providing datasets of harassing messages [38, 49] or improved detection models [40, 65, 71]. However, the utility of automatic harassment detection can be limited as methods of harassment continually evolve with the development of social technologies [21, 75]. There is also the risk of incorrect labeling, often due to lack of context [91, 99] or disproportionate moderation of content from marginalized groups [41]. In addition, the reactive nature of such platform-led interventions means that there is often a delay until they are processed by human moderators, during which harm may persist [82, 94]. Roberts emphasizes the limitations of commercial content moderation in that they "always attract...an existence as a generator of the negative content" [81]. This is further bottlenecked by the sheer volume, lack of guidance, and conflict of commercial interest. Thus, the importance of user-led bottom-up interventions for harassment response is increasingly emphasized.

There has also been an emphasis on designs to prevent harassment incidents before they happen. Previous work has explored methods such as accentuating the sense of personal responsibility [24], promoting empathy towards the harassment victim [98], and using negative interface cues to influence perception [58] to reduce aggression while encouraging supportive action from the bystanders. At a lower level, systems such as Recast [105] aims to reduce harassment by detecting toxic language and intention, discouraging users from posting harmful messages. This is especially crucial in the context of networked harassment, as preventing amplification and reducing the scale of the incident is can significantly reduce its negative effects.

Where systematic efforts fall short, leveraging communities has shown to be an effective method to combat harassment. Many platforms, such as Reddit and Twitch, employ volunteer moderators that independently and flexibly adapt to the moderation needs of their community [91]. Outside of self-moderated contexts, others have explored facilitating bystander intervention as an effective method of limiting harassment as well as empowering the victim of harassment [15, 16, 98, 104]. Tools for individual moderation, such as blocking, could also be utilized for harassment prevention. Community-created collaborative blocklists, such as BlockTogether and Good Game Auto Blocker are an example of a collaborative community effort to protect individuals from potential harassers [39, 46]. In a more personal level, systems such as Squadbox utilizes a user's friend groups as a moderation tool [64] , while some communities on Twitch has leveraged their audiences to respond to hate raids [69]. These methods allow for more personalized and intimate methods of protection from harassment.

However, despite the extent of research on content moderation and harassment intervention, we emphasize the importance of measures that prioritize the individuals involved. Focusing on the victims' experiences is crucial in restoring trust towards the platforms and supporting them to overcome the experiences [9, 90, 96]. It is also important to understand that there is no 'one-size-fits-all' solution to online harassment [88] - every case is unique, and as harassment techniques evolve, proposed solutions must remain flexible as well. Thus, we focus more on exploring methods for users to independently and effectively combat harassment as an individual. We expand upon this line of work by emphasizing design choices that empower the individual to quickly counteract harassment in ways that meet their personal needs.

2.3 Focusing on the Needs of Harassment Victims

As noted above, definitions and perceptions of harassment often differ by the user, and the ways through which the users want to establish justice also vary depending on the cultural and social contexts of the involved parties. Platforms' responses to online harassment usually focus more on punishing the perpetrators of harassment rather than protecting or reassuring the harassed user [68, 75]. In addition to preventing further incidents, providing closure and validation incidents has also been explored as a method to emotionally support victims of harassment. Platforms such as Heartmob [9] and Trollbusters [29] tackle this by providing direct and indirect support to victims, as well as a safe space to share their experiences, facilitating recovery.

Schoenebeck et al. have previously explored the potential of utilizing alternative justice theories, such as restorative justice, to understand different needs surrounding reparations to online harassment [88]. From a systems perspective, Sultana et al. explore the concept of shame-based justice in *Unmochon*, where public shaming is used to mitigate and prevent sexual harassment in Bangladeshi women's social media experiences [96]. We expand upon this scholarship and aim to provide a novel perspective surrounding anti-harassment systems based on the experiences of South Korean Twitter users. We also note that anti-harassment systems should allow for flexibility and variability according to the needs and situations of the individual.

This work also builds upon feminist methodologies in HCI, which focuses on emphasizing the lived experiences of individuals and exploring how these can benefit the design process [3]. An example of this is the discussion of promoting consentful technologies in the online space. Im et al. have previously discussed the potential of applying an affirmative consent [59] framework to the design of social media systems [43], and how users may be protected from interactions that they are uncomfortable with. In this perspective, networked online harassment is a violation of consent where users are bombarded with responses at a scale they did not anticipate nor agreed to [46]. Thus, we argue that the ability to more specifically control one's social media audiences can allow harassment victims to reclaim their agency and provide the ability to resist.

3 DESIGN ITERATION: DESIGN WORKSHOP

We designed an initial prototype of the system with the goal of promoting civil discussion, while making it easier to protect the original poster from the unexpected repercussions of amplification. We initially focused on the 'unintentional' aspect of networked harassment, or that not all perpetrators of networked harassment will be acting from malicious intent [43, 52, 66]. In this perspective, some people become perpetrators because they did not understand the potential repercussions of their actions prior to commenting, unexpectedly or unintentionally contributing to networked harassment through amplification. Assuming such cases with non-harmful intent, our initial design provided users with a way to initiate conversation and discussion without amplifying the original post.

We iterated the system design through three workshop sessions with 11 Twitter users who had experiences surrounding networked harassment and calling out. We aimed to verify our prototype and gather feedback while collecting potential use case scenarios and gaining a better understanding of methods and key factors in preventing networked harassment. We describe our methodology and insights from the workshop in the following section.

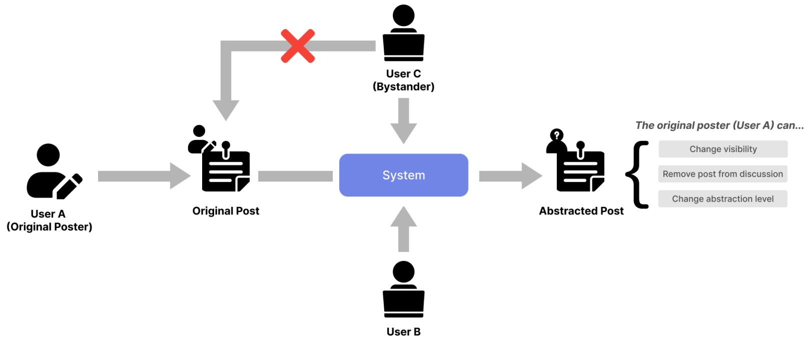


Fig. 1. Overall User Flow of Workshop Prototype

3.1 Initial Design Prototype

We first conceptualized our system with the concept of 'separating' the post, and the subsequent reactions to it, from the person who posted it. We refer to this feature as 'abstraction', allowing the post author and post to be disconnected by involving a mediator system that hosts the discussion. In this case, an 'abstracted' post refers to content that does not contain references to the original post or OP's account, and cannot be directly traced back to the OP. We extended beyond simply anonymizing posts by also suggesting a text obfuscation process, where the text would be replaced with non-traceable altered version of the original post. This prevents other users from re-tracing the original post and poster through an exact text search.

Our idea of abstraction was inspired by the 'orphaning' feature in Archive of Our Own[1], a popular fanfiction archive website. Orphaning is when an author of an online work decides to remove their association with it, providing a level of control to the authors' online presence while allowing the work itself to persist to preserve the community history and discourse surrounding it [30]. Similarly, we aimed to enable and encourage discussion over a topic, considering the positive overall value of discourse it may bring, while allowing the original poster to maintain a level of control over the involvement of their personal identity and account.

When using the system, a user who wants to comment on another user's post is provided with the option to 'Start Discussion', in addition to the original retweet and quote tweet functions. The user is given the ability to add their opinion and post it, as they would with the quote tweet function. This will generate an abstracted, mediated version of the tweet, which is hosted in a separate database. This instance of the abstracted tweet aggregates all the individual discussion nodes pertaining to the original Tweet. The mediated version does not save any identifying information pertaining to the OP, such as profile information, while the text is obfuscated to prevent malicious users from searching up the post content to find and harass the user. The overall user flow is depicted in Figure 1.

t

3.2 Participants

Through a public Twitter post, we recruited 11 South Korean Twitter users who reported to have had experience surrounding networked online harassment, either as victims, perpetrators, or bystanders. We note that our definition of 'perpetrators' also included those who had unintentionally or

[1] https://archiveofourown.org/

unknowingly participated in networked harassment through calling out behaviors [52]. Participants were asked to select all categories that applied to them out of the three, as well as provide a short description of their experience when signing up.

We verified the validity of the reported experiences through keyword searches based on the description as well as references provided by the participants when available. When we were unable to obtain specific evidence – either due to the time passed since the incident, the deletion of relevant evidence, or due to the abundance of similar incidents – we reviewed the participants' comments for internal consistency and credibility. We acknowledge that there is an imbalance of experience types in our participants, skewing toward victim experiences, as well as showing significant overlap between the experience types. This can be attributed to the fact that victims of networked harassment often have broader definitions of what constitutes harassment [52], which may have altered their perceptions of their previous actions. Also, due to the distributed nature of networked harassment and the reduced personal impact of the incident, perpetrators of networked harassment may not self-identify as perpetrators [45, 46, 52, 66]. The participant demographics are detailed in Table 1.

3.3 Methods

Participants were divided into groups of 3 to 4 people and participated in a collaborative design workshop through a Zoom video call. The groups were built primarily based on the participants' availability, and to include a variety of experiences regarding the type and subject of the incident. We provided a low-fi prototype of the system designed through Figma and used screenshots of the UI, simulated user workflows, and potential use case scenarios to illustrate the system features.

To ensure the participants' safety and comfort, all participants communicated through pseudonyms assigned by the researchers. Referring to each other through pseudonyms is a commonly-observed practice in Twitter [76], as well as South Korean online communities [57], and we aimed to provide a similar experience through this design. We also didn't require the use of video, and only used audio and text during the workshop to prevent participants from unintentionally identifying themselves or creating additional social pressure. This anonymized, verbal-centric environment was used to emulate the experience of Twitter and provide a smoother transition in discussing their own experiences on the platform. Prior to the workshop, participants were advised to defer from disclosing sensitive or personally identifiable information, and to only share information they were comfortable with.

The workshop consisted of three sections. First, we asked the participants to share their experiences and opinions surrounding public criticism and online harassment. Following that, we introduced our system prototype, intended user flow, and design rationale behind the features. Participants were then asked to collaboratively suggest alternative features that could be added, while imagining potential use case scenarios that might utilize our system. We encouraged participants to build upon others' ideas and to point out the limitations or potential negative repercussions of the system and the generated ideas. Finally, we moved on to a group interview session where the participants provided feedback on the system design and discussed what elements should be considered when building social media systems that combat online harassment.

We performed qualitative coding on the workshop transcripts, with two of the authors individually identifying codes from the initial transcripts which were combined to create a codebook. The first author then used this codebook to re-code the transcripts focusing on the specific feature suggestions made by the participants.

Table 1. Participatory Design Workshop Participant Demographics.

ID	Gender	Group	Age	# of Accounts	Networked Harassment Experience		
					Perpetrator	Victim	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

3.4 Workshop Results

Based on the discussions from the workshop, we organize the findings and design implications pertaining to online and networked harassment. During the group interview session, many participants shared the sentiment that networked harassment was perpetuated by the nature and design of Twitter as a social media platform, especially due to the emphasis on amplification features. Many agreed with the belief that criticism and calling out easily turns into harassment when the critical opinion overwhelms the supportive as a result of networked amplification. Participants also noted that networked harassment happened as people wanted to "*feel safe among the masses*." (W2). These comments imply that some Twitter users thought of online calling out as being a mob mentality behavior rather than stemming from genuine concern or critique – and thus, closer to networked harassment than valid criticism. Below, we summarize the specific design goals and features suggested by the participants.

3.4.1 Accurately Reflect the Context and Intent of the Original Poster. One of the major comments from the participants was that, if abstracted, the content should fully and accurately transfer the OP's intent, tone, and context (W6, W4, W10). 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 meant that diverse nuances and contexts are packed into a small amount of text, which could be easily misrepresented when the content is altered even slightly, and especially when it spreads beyond the initial intended audience. Thus, original poster may have to take responsibility for comments beyond what they had actually expressed.

Participants also noted that users should retain agency over how their post is expressed and how others might perceive it, while preventing the potential misuse of such features. For example, discussion on features allowing users to be able to edit the post directly were 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 edits or clarifications to misleading statements. 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."*

3.4.2 Allow Specific and Granular Control Over One's Post.

In relation to the above point, participants voiced that the original poster should maintain a level of agency in the conversation, both in discussion and moderation. 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). Thus, they emphasized the importance of having personalized and granular control of various settings so that each user could alter them as they see fit. Our participants mentioned that granular notification settings could be an example of this, as it is one of the major factors that distinguished criticism from harassment was scale. Participants noted that the large influx of notifications often caused social pressure, as well as a feeling of helplessness. Thus, many participants suggested that the system provide specific controls for notifications (W5, W10).

3.4.3 Prevent Amplification of Posts.

Participants also noted that it is important to reduce the spread of criticism to ensure that the negative effects of calling out and harassment could be mitigated (W2, W8, W10). This aligns with the findings from previous work stating that the scale of calling out is a large determining factor of online harassment [52]. W10 articulated this, saying: *"I think the problem is receiving the 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. Some participants suggested features to allow the OP to nip the harassment at the first signs, as well as informing them of the potential scope of amplification. 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.

3.4.4 Emphasize Responses Encouraging Constructive Discussion.

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

3.5 Design Goals

Based on the 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. We identified key features suggested from each theme of the workshop results, from which we then identified the users' key needs, organized into our design goals. The mapping between the themes, suggested features, and the design goals are illustrated in Figure 2.

> **D1. Provide victims the ability to distance themselves from open audience spaces** to prevent interactions with harassers (*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*).

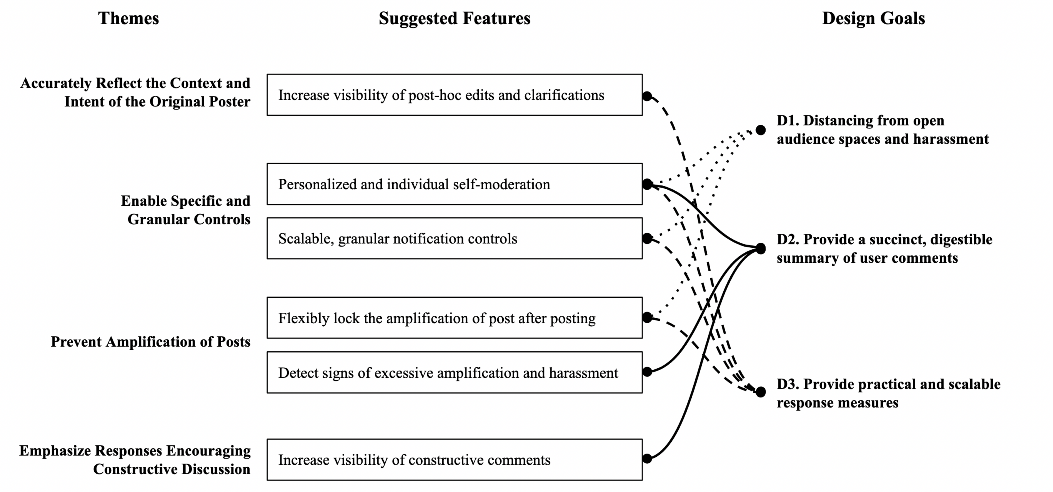

Themes
Suggested Features
Design Goals
Accurately Reflect the Context and Intent of the Original Poster
Increase visibility of post-hoc edits and clarifications
Enable Specific and Granular Controls
Personalized and individual self-moderation
Scalable, granular notification controls
Prevent Amplification of Posts
Flexibly lock the amplification of post after posting
Detect signs of excessive amplification and harassment
Emphasize Responses Encouraging Constructive Discussion
Increase visibility of constructive comments
D1. Distancing from open audience spaces and harassment
D2. Provide a succinct, digestible summary of user comments
D3. Provide practical and scalable response measures

Fig. 2. Mapping of Design Goals with Themes and Suggested Features from the workshop

4 RE:SPECT

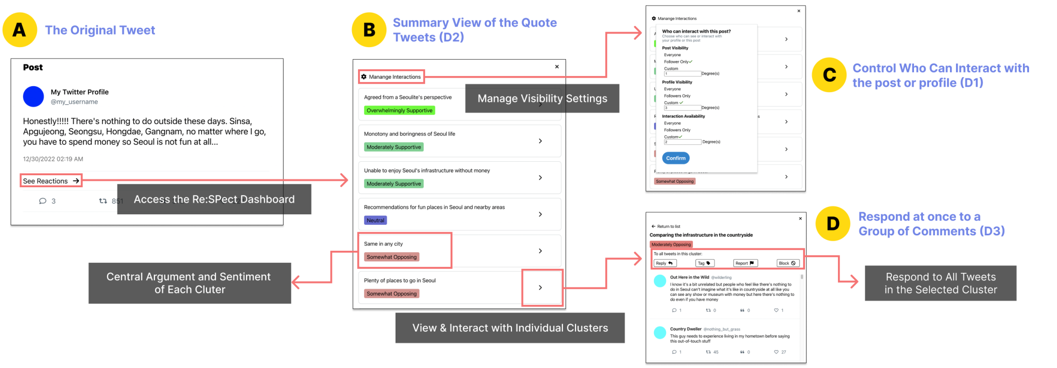
Based on our design goals, we revised our original concept to design *Re:SPect*, a system that provides active and scalable response measures for victims of networked online harassment and calling out. In Re:SPect, we focused on the perspective of protecting the victim, as well as providing practical response measures for them to respond against online harassment. 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.

As we focused on the issue of networked harassment on Twitter, the system design was built with relevant Twitter features in mind, and we envision Re:SPect to be potentially deployed as a plugin or extension to Twitter or similar platforms. However, due to the fact that our design of Re:SPect could not be completely implemented upon Twitter, we developed a testable prototype where Re:SPect's features were integrated into Twitter's user flow. 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.

4.1 System Features

When a user views their post, they are provided with the option to 'See Reactions'(Fig. 3-A). Clicking on this button leads them to the dashboard interface that organizes the information about the post as well as its responses (Fig. 3-B).

4.1.1 DG1: Control Who Can See or Interact with the Post. In the dashboard, the OP can access the 'Manage Interactions' tab to control who can see or interact with their post (Fig. 3-C). This controls how the post is displayed to others in three different categories: the visibility of the post, the visibility of the user profile, and who can interact with the post. When a user is outside of the distance conditions set by the OP, 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 OP (Fig. 4-Right).



Post

My Twitter Profile @my_username

Honestly!!!!! There's nothing to do outside these days. Sinsa, Apgujeong, Seongsu, Hongdae, Gangnam, no matter where I go, you have to spend money so Seoul is not fun at all...

12/30/2022 02:19 AM

See Reactions →

Manange Interactions

- Agreed from a Seoulite's perspective — Overwhelmingly Supportive
- Monotony and boringness of Seoul life — Moderately Supportive
- Unable to enjoy Seoul's infrastructure without money — Moderately Supportive
- Recommendations for fun places in Seoul and nearby areas — Neutral
- Same in any city — Somewhat Opposing
- Plenty of places to go in Seoul — Somewhat Opposing

Who can interact with this post?
Choose who can see or interact with your profile or this post

Post Visibility: Everyone / Follower Only ✓ / Custom [1] Degree(s)

Profile Visibility: Everyone / Followers Only / Custom ✓ [3] Degree(s)

Interaction Availability: Everyone / Followers Only / Custom ✓ [2] Degree(s)

Confirm

← Return to list

Comparing the infrastructure in the countryside — Moderately Opposing

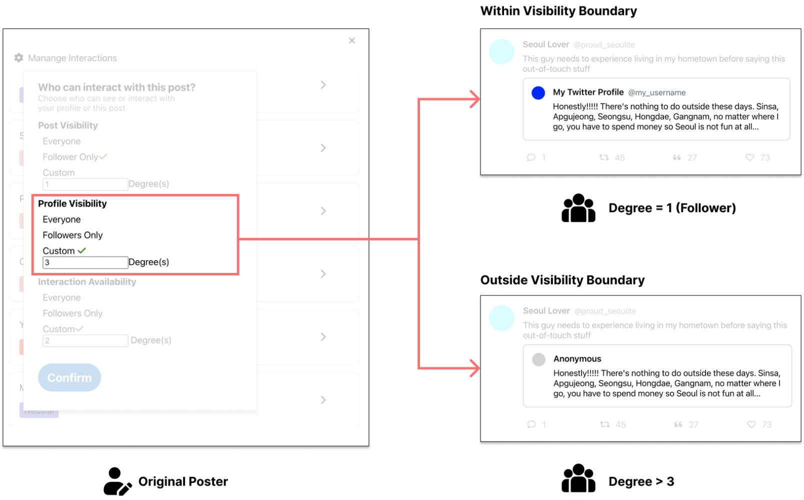
To all tweets in this cluster: Reply / Tag / Report / Block

Out Here in the Wild @wilderling
I know it's a bit unrelated but people who feel like there's nothing to do in Seoul can't imagine what it's like in countryside at all like you can see any show or museum with money but here there's nothing to do even if you have money
1 0 0 1

Country Dweller @nothing_but_grass
This guy needs to experience living in my hometown before saying this out-of-touch stuff
1 45 0 27

A
The Original Tweet
3
851
Access the Re:SPect Dashboard
B
Summary View of the Quote Tweets (D2)
Manage Visibility Settings
Central Argument and Sentiment of Each Cluter
View & Interact with Individual Clusters
C
Control Who Can Interact with the post or profile (D1)
D
Respond at once to a Group of Comments (D3)
Respond to All Tweets in the Selected Cluster

Fig. 3. Overview of the User flow of Re:SPect. (A) The user accesses the Re:SPect dashboard through the 'See Reactions' button on the tweet. (B) The dashboard view of Re:SPect shows the clustered responses, as well as the central argument and sentiment of each cluster. (C) The user can access the visibility settings tab to control the post and profile visibility and interactions boundary. (D) In the detail view of each cluster, the post owner could perform actions en masse to all the responses in the cluster.



Manange Interactions

Who can interact with this post?
Choose who can see or interact with your profile or this post

Post Visibility: Everyone / Follower Only ✓ / Custom [1] Degree(s)

Profile Visibility: Everyone / Followers Only / Custom ✓ [3] Degree(s)

Interaction Availability: Everyone / Followers Only / Custom ✓ [2] Degree(s)

Confirm

Within Visibility Boundary

Seoul Lover @proud_seoulite
This guy needs to experience living in my hometown before saying this out-of-touch stuff

My Twitter Profile @my_username
Honestly!!!!! There's nothing to do outside these days. Sinsa, Apgujeong, Seongsu, Hongdae, Gangnam, no matter where I go, you have to spend money so Seoul is not fun at all...

1 45 27 73

Outside Visibility Boundary

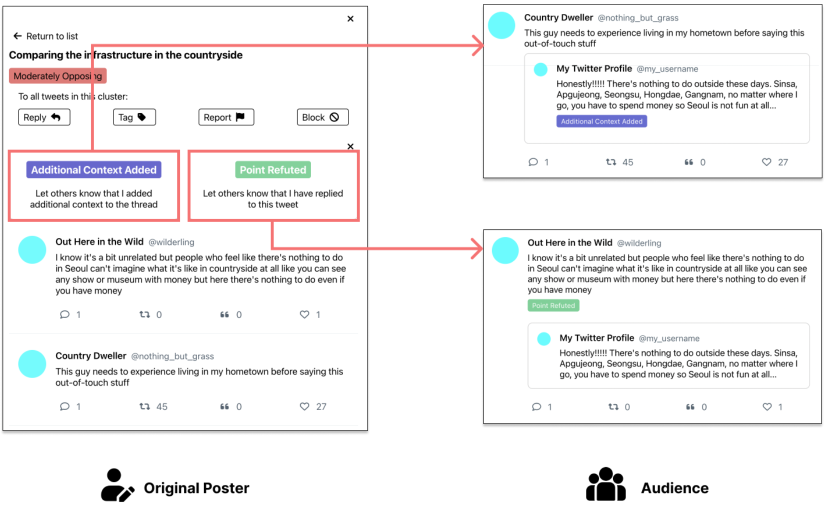
Seoul Lover @proud_seoulite
This guy needs to experience living in my hometown before saying this out-of-touch stuff

Anonymous
Honestly!!!!! There's nothing to do outside these days. Sinsa, Apgujeong, Seongsu, Hongdae, Gangnam, no matter where I go, you have to spend money so Seoul is not fun at all...

1 45 27 73

Fig. 4. 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 OP's account information. (Right) When the viewer is within the profile visibility boundary.

Each visibility category has three options: (i) **Everyone**, meaning that everyone on the social network can access the post, (ii) **Followers-only**, referring to the immediate follower network of the user, and finally (iii) **Custom**, where the OP can personally define how far 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 OP. We note that higher-level visibility settings such as post or profile visibility are also automatically applied to the lower-level settings. For example, if


×
Return to list
Comparing the infrastructure in the countryside
Moderately Opposing
To all tweets in this cluster:
Reply
Tag
Report
Block
×
Additional Context Added
Let others know that I added additional context to the thread
Point Refuted
Let others know that I have replied to this tweet
Original Poster
Audience
Additional Context Added
Point Refuted

Out Here in the Wild @wilderling
I know it's a bit unrelated but people who feel like there's nothing to do in Seoul can't imagine what it's like in countryside at all like you can see any show or museum with money but here there's nothing to do even if you have money
1 0 0 1

Country Dweller @nothing_but_grass
This guy needs to experience living in my hometown before saying this out-of-touch stuff
1 45 0 27

Country Dweller @nothing_but_grass
This guy needs to experience living in my hometown before saying this out-of-touch stuff

My Twitter Profile @my_username
Honestly!!!!! There's nothing to do outside these days. Sinsa, Apgujeong, Seongsu, Hongdae, Gangnam, no matter where I go, you have to spend money so Seoul is not fun at all...
1 45 0 27

Out Here in the Wild @wilderling
I know it's a bit unrelated but people who feel like there's nothing to do in Seoul can't imagine what it's like in countryside at all like you can see any show or museum with money but here there's nothing to do even if you have money

My Twitter Profile @my_username
Honestly!!!!! There's nothing to do outside these days. Sinsa, Apgujeong, Seongsu, Hongdae, Gangnam, no matter where I go, you have to spend money so Seoul is not fun at all...
1 0 0 1

Fig. 5. 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. (Right, Top) An 'Additional Context Added' flag is added to the original tweet. (Right, Bottom) A 'Point Refuted' flag is added to the quote tweet.

one is outside of the post visibility boundary, they are naturally also outside of the profile visibility boundary.

4.1.2 DG2: Summarization of Responses. Re:SPect provides a digestible and summarized view of the responses on the original post. This provides users with the ability to have a more balanced perception of the overall opinion distribution. This can also have 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 and 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.

4.1.3 DG3: Responding to Comments at Scale. Users can also choose to respond to a specific cluster of comments. In the Response Cluster Tab, users can select the type of response measure to take (Fig. 3-D). The system automatically batch-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 OP to add additional context to the original post while controlling how they are perceived by the networked audience.

The OP may also choose to use post flags to respond to a cluster. There are two types of flags that the OP can utilize: 'Additional Context Added' and 'Point Refuted' (Figure 5). 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 there is additional context to the conversation.

5 METHODS

We decided against conducting a real-life field study as it is not possible to simulate realistic harassment situations, as well as concerns that Re:SPect could cause negative repercussions to the harassment situation or to the users. 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 speculative study design follows the precedent of anti-harassment systems research such as Squadbox [64] and Unmochon [96].

5.1 Study Setup

To observe how participants may use Re:SPect in real harassment scenarios, we provided them a set of 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. To ensure that our scenarios reflected real-life conditions as closely as possible, we used the data of an existing tweet that had recently been the target of a networked calling-out, harassment incident. The selected tweet was a casual complaint regarding the lack of exciting activities to do in Seoul, South Korea. The selection was based on the perceived relevance of the situation (Can the participants relate to the tweet content?) as well as the scale of the responses. The translated tweet content is depicted in Figure 3-A. The majority of our participants expressed explicit agreement with the content of the tweet, and the few who expressed disagreement or neutrality stated they could understand where the OP was coming from and had no difficulties assuming the tweet's position. We used the Twitter API to crawl the responses to the tweet to ensure the situation felt realistic. To minimize unintended harm and discomfort to the participants, we filtered out quote tweets containing potential triggers or personally identifiable information.

Based on the collected data, we modified the data to create two scenarios, preemptive and reactive, representing different points in time in the progression of networked harassment. The reactive scenario depicted the time frame after the harassment had happened: the entirety of the quote tweet dataset (392 of the accessible public quote tweets) was applied to this scenario. In comparison, the preemptive scenario represented a situation where the tweet has reached a networked audience, but at a smaller and individually manageable scale. We filtered out 15 quote tweets from the original dataset, focusing on the ones with earlier posting times (representing the initial response) and with less aggressive language (representing a 'tamer' incident of networked response). The preemptive scenario was constructed based on Kim et al.'s observations of factors that distinguish between online harassment and valid criticism [52], including the scale and aggression level of the responses. 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.

We also note that, for the purposes of the current study, we implemented the summarization feature (**D2**) through a manual clustering of tweets by the authors. While our ideal implementation of the summarization feature would be for the clustering process to be integrated through automated methods, our survey of existing NLP models proved lacking for the purposes of our study. Since we are not conducting a formal evaluation of our technical pipeline, but rather seeking a proof-of-concept for our system design, we assumed an ideal clustering situation as in the results of our manual clustering.

5.2 Interviews

The interviews were conducted through Zoom video calls, lasting between 85 and 119 minutes. The session began with a preliminary observation of the 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 system usage session where participants used Re:SPect to demonstrate their potential responses to online harassment. The system 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 to social media platforms through a semi-structured interview. Participants were each paid 30,000 KRW (approx. 23.6 USD).

The interview results were transcribed by the authors for analysis. We conducted thematic analysis through an inductive approach, developing and validating the codes through multiple iterations [27]. The first and second authors individually performed line-by-line open coding on six of the interviews to generate preliminary codes, then combined the results to create an initial codebook. 34 codes were derived from the initial analysis. Based on the initial codebook, the second author iterated over the remaining interviews, developing 40 codes in total. The final codebook was then validated and used for re-coding by the first author. After the final iteration, we developed similar codes into themes, organizing them first by the core factors on dealing with networked harassment, perspectives on Re:SPect's features, and the potential use cases and impact of the features general harassment situations. Quotes have been translated and edited for clarity by the authors.

5.3 Participants

We interviewed a total of 18 South Korean Twitter users, who were recruited through a public Twitter post, as well as personal recommendations from previous participants. Four participants in total were invited through snowball sampling, all through the recommendation of different participants. The condition for participation was Twitter users who have either personally experienced (*direct experience*) networked online harassment, or have witnessed and speculated about tactics to mitigate the negative effects of networked online harassment (*indirect experience*). The participant demographics are detailed in Table 2.

We note that, as with the case of the workshop, the participants were predominantly female, with only one participant identifying as a cisgender male. This can be attributed to the fact that Twitter is considered a social media where the political views are relatively progressive in South Korea, including female-centric social issues such as feminism [42, 50]. This may mean that the Twitter community in South Korea is already female-dominated. From a different perspective, this may be attributed to the fact that marginalized groups such as sexual and gender minorities, racial minorities, etc. are often more at risk of online harassment in social media [20, 26, 31, 60]. Previous work has also shown that women are also more commonly exposed to gender-based harassment [20, 74, 101]. Thus, we can assume that women and LGBTQ+ individuals may be more conscious of the potential of online harassment, which may have affected our participant distribution.

Table 2. Re:SPect User Study Participant Demographics.

ID	Gender	Cisgender/ Transgender	Age	# of Accounts	Experience Type		Purpose of Twitter Use
					Direct	Indirect	
P1	F	Cisgender	24	4		O	Kpop Fandom, Anime fandom, Daily life sharing
P2	M	Cisgender	29	2		O	Socializing with offline friends, Exercise sharing
P3	F	Cisgender	24	3		O	Socializing with offline friends, Personal archiving
P4	F	Cisgender	24	3		O	Research networking, Stationery & tea collecting
P5	F	Cisgender	30	3		O	Study, Daily life sharing
P6	F	Cisgender	24	6	O		Art sharing, Daily life sharing, Socializing with offline friends
P7	F	Cisgender	23	5	O		Kpop fandom, Misc. fandom
P8	F	Cisgender	28	3		O	Anime fandom, Combating harassment
P9	F	Cisgender	22	3	O		Kpop-related art sharing
P10	F	Cisgender	26	3	O		Queer community, Disability community
P11	F	Cisgender	23	3	O		Daily life sharing
P12	Does not wish to answer		24	3		O	Political discourse, Daily life sharing
P13	F	Cisgender	27	2	O		News sharing
P14	Unknown	Transgender	26	2	O		Daily life sharing
P15	F	Cisgender	20	1	O		Daily life sharing
P16	Does not identify		20	6	O		Game fandom, Art sharing, Online market
P17	F	Cisgender	26	3	O		Daily life sharing, Trading used items
P18	F	Cisgender	25	2	O		Kpop fandom

6 RESULTS

In general, participants reacted favorably towards the concept and implementation of Re:SPect. 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 encourage 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 design goals (*Protect*, *Summarize*, and *Response*) were met.

Here, we organize the common themes that emerged from the interviews, focusing on the central factors in mitigating networked online harassment and the perceived use cases of each feature. We identified three major factors that users considered important in preventing and mitigating networked harassment: preemptively reducing harassment by reducing amplification and exposure, shielding users against the immediate negative effects of harassment, and providing scalable response measures. We evaluate the feasibility of Re:SPect in achieving these goals, while also discussing its limitations and identifying potential aspects to improve Re:SPect.

6.1 Preventing Networked Harassment by Reducing Amplification

Participants had generally favorable expectations for Re:SPect in preventing the occurrence of online harassment. This effect was observed twofold, as participants noted that Re:SPect would be effective in both stopping harassment from happening at all, and also in preventing the further amplification of harassment. This was also connected to the importance of timely and appropriate responses. Many participants pointed out that in networked harassment, if they missed the 'golden time' of stopping harassment, it would spiral out of control to a level that can't be managed. 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

6.1.1 Protection from Exposure to Malicious Actors. Many participants also noted that the 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', which 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.*"

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.*"

6.1.2 Reducing Possibility of Secondary Harassment. 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 avoiding it altogether, or just waiting. - P7

Several participants touted Re:SPect's support in staunching the spread of misinformation or misinterpretation early on through signposting (P4, P5, P10, P11, P13). Though the additional content might not be visible on the quote tweets, participants felt that simply indicating the existence of further context would deter observers from heedlessly following the quoter's opinion (P4, P10, P13, P18) and motivating them to look into the additional information (P8, P11, P12).

> I think the [post flagging feature] will be useful for both the users who made QTs and those who read them. It makes responders reflect on their words and reduces the chances of secondary harassment from observers. - P18

6.2 Reducing the Negative Emotional Impact of Networked 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 or a safety net from potential harm emerged from several interviews (P2, P3, P4, P6, P12, P17). Several participants (P3, P6, P8, P10, P15) noted that the sheer existence of anti-harassment tactics will make them fear harassment a lot less, providing them with 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.*" In this section, we delve into more detail on why users felt safer, as well as which features of Re:SPect contributed to protecting users from the negative psychological impacts of harassment.

6.2.1 Allowing for More Accurate and Efficient Information Processing. Participants mentioned that the summarization feature allowed them to perceive the responses and opinions more clearly and objectively. In many cases, participants noted that they would focus on malicious or negative comments even when there were positive or supportive comments, such as when P1 said 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. Many participants reported that they would gravitate toward the positive clusters instead of the negative ones. 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 given the option, they would opt to ignore the negative comments completely, focusing only on the positive comments and not reading the contents of the negative comments.

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 what is happening in the quote tweets in a networked harassment situation, due to the scale of responses being more than they could process. Reading through each individual comment and mentally processing and compartmentalizing each opinion caused fatigue to the participants, leading to them avoiding checking the responses altogether. However, with the summary view, participants said that they would feel less burdened and that they will check the responses more often as a result (P4, P5, P17), eventually being more actively engaged in the conversation.

6.2.2 Protection from Immediate Exposure to Negative Responses. The response summarization feature was also lauded by many participants in that it protected them from being exposed to negative comments before they were prepared. Many participants noted that on Twitter, there is no way of knowing what kind of responses you may have before actually checking them. Such uncertainty often led to anxiety about viewing the responses, especially when they were at an unexpected scale. 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) will often be simple vitriol or aggressive opinions that are "*not even worth responding to.*" (P2). As they perceived these opinions and clusters to have minimal communicative or informational value, they would simply forego reading and interacting with such negative opinions at all. Participants such as P5, P10, and P17, noted that people would act more carefully to not be categorized in extreme categories, pointing out that valid critiques are still harassing when the language used is overly aggressive. However, several participants did note 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).

6.3 Taking the Initiative to Respond

Several participants pointed out that the central problem of 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). Thus, by providing response measures that could operate at scale, we were able to provide users with a sense of security that they will be able to react and respond even in situations of harassment. This sense of self-efficacy and agency encouraged participants to use diverse methods to actively speak up about their perspectives.

6.3.1 Providing Methods for Effective Clarification. The post flagging feature was noted as an important feature that allowed users the opportunity to actively try to redeem themselves in the face of misunderstandings. This allowed participants to interject the conversation with new information to resume their control over the conversation. The participants felt that their clarifications or rebuttals went ignored on Twitter, especially during the peak of the spread and vitriol. This would result in a barrage of identical arguments flooding and further burying their additional tweets. The problem is exacerbated by third parties who see the quote tweet first and formulate their opinions under its influence.

> 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

Thus, most participants appreciated Re:SPect's ability to always bring their content front and center. 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, 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.

6.3.2 Encouraging the Use of Active Responses. 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. P14 noted how the features of Re:SPect helped "*my voice to be more evenly matched to that of the harassing group.*", which led to increased self-efficacy in their responses. 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 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

6.4 Additional Use Cases and Potential Drawbacks of Re:SPect

Adding to this, our participants pointed out several potential use cases of Re:SPect that we had not designed for, but could occur as a result of the features that we had presented. This included both positive use cases, such as more effectively spreading information and prompting healthy debate, as well as negative ones such as developing novel harassment pathways and avoiding responsibility. We detail them in the following sections.

6.4.1 User-led Spread of Information. Our findings from the workshop suggested that a key damaging aspect of networked harassment is how fast information – some of which may be incorrect or misinterpreted – is spread. In Re:SPect, the visibility control features and context post flags were implemented to control both the range of the spread and how it is interpreted, respectively. However, some of our participants noted that the system can actually also be used to disseminate information faster for their own needs (P4). For example, they can add updates to existing tweets (e.g., notifying people of location changes for an event) through the mass-reply function to quickly notify many individuals at once. The user can also choose and customize the spread to designated groups, not just for the sake of responding to ongoing harassment but also for more practical reasons such as appealing to different interest groups or respecting the content preferences of their followers (P17).

6.4.2 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 room 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 quote-tweeting 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

6.4.3 Harder to Take Responsibility for 'Wrong' Actions. 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.

6.4.4 Re:SPect Could Exacerbate Harassment. Some participants voiced concerns over how the features of Re:SPect could be abused to heighten the harassment. Similarly to the above point, P2, P4, and P8 noted that people who harass others might attempt to hide their profiles and avoid the consequences. On the other hand, P5 and P8 noted the audience may feel even less restrained when harassing others with anonymized profiles due to the depersonalization of the author. Some participants were also wary of using the response measures for fear of inflaming the discourse and attracting even more unwanted attention. P11 explained, "*If no one is backing off, it just becomes a fighting ring.*"

> Even if you haven't done anything wrong, if you overreact, you can become the bad guy [in their eyes]. (...) If you respond more actively, it might actually just aggravate the situation. - P3

A couple of participants noted that the summarization feature brought negative viewpoints into the spotlight, which could cause stress and anxiety (P5, P7). They also noted that some users, especially those with more fragile mental states, might end up catastrophizing the situation by only focusing on the negative comments or even 'doomscrolling' with the easily-accessible summary of negative comments.

6.5 Usability Concerns

In addition to the potential use cases, some participants provided interface-level usability concerns and elements for improvement regarding the design of Re:SPect. P17 and P18 noted that the system's UI depth was generally large and that the nested views requiring many clicks to access the settings could be too complex. This led to the perceived complexity of the system, and some participants thought that it would take some training before they could fully utilize the functions. In this line of thought, P1 and P5 also noted that the visibility settings should be on a separate view, as in the case of the reply settings of Twitter, and not included in the dashboard. We note that the tools we provide must remain accessible to ensure that users are able to employ the response measures in real harassment situations. Thus, the learning curve and complexity of Re:SPect is something we should take care to improve in future iterations should such features be applied to real-life scenarios.

7 DISCUSSION

The results of our study suggest that Re:SPect was able to promote more active and scalable responses to networked online harassment, while relegating a sense of control back to the victims. Many participants praised the diversity of solutions that Re:SPect provided, as it allowed them to pick and choose the solution that worked best for them. This aligns with previous findings which state that individuals with different experiences have different expectations and preferences for online justice [88]. Several participants used a different combination of features while aiming to achieve the same effect, which suggests that Re:SPect could potentially be applied to diverse harassment scenarios beyond what was explored in the current study. However, while Re:SPect's features were unanimously praised for their potential to protect users against networked online harassment, some did note that they could also be used to exacerbate echo chambers, refuse to heed valid criticism, and abuse anonymity. In this section, we build upon these insights and discuss how this work situates within the larger body of online discourse literature, as well as key factors involved with improving the users' agency in the face of networked harassment.

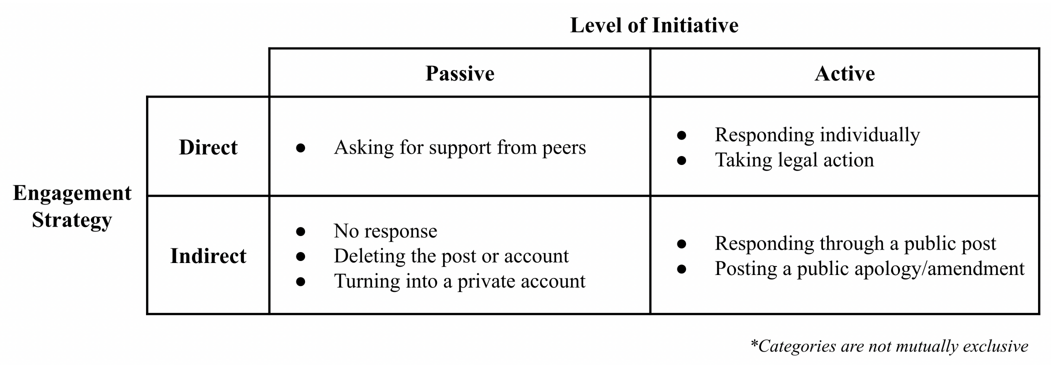
7.1 Promoting Safer and Healthier Online Discourse

Online harassment in social media is considered a significant threat not just to the individual, but also to the platform itself. Fear of conflict or harassment discourages individuals from speaking up or expressing their thoughts online [52, 60, 68], reducing the amount of engagement on platforms. On the other hand, online discourse is often where many conflicts and antisocial behavior starts. Thus, discussing how we might reduce online harassment on social media is inevitably connected to promoting healthier online communication. While Re:SPect focuses primarily on reducing the negative repercussions of harassment on the individual, our participants suggested that its features could have significant implications for how discourse is shaped on Twitter.

Re:SPect achieves this by allowing the victims of networked harassment to resume control of the discussion. Through features such as post flags and interaction boundary settings, the users can act as a moderator of their own posts to ensure that the discussion stays within an acceptable boundary. This builds upon Im et al.'s work on affirmative consent, who introduced the idea of consentful social media systems that allow users to freely and flexibly exert their control over their posts [43]. Online spaces will always be plagued by "a few bad apples" – malicious actors whose primary intention is to hurt others [77]. Rather than focusing on reconstituting these individuals, Re:SPect puts the consent of the individual first by determining who and how they see the content through the visibility controls and flagging functions. As the conversation evolves, the owner of the post can reel back deviant behavior and signal their boundaries, informing others on how they should engage.

The features can also encourage meaningful and constructive discussions. Current forms of online 'discussions' are often fatigue-inducing; users become overwhelmed by the anticipation of diverse interpretations and reactions to their words and feared being wrongly accused without even reading the content. Features such as the flagging system and mass interaction functions of Re:SPect could enable users to resolve misunderstandings effectively and efficiently. This ensures that the audience and the author are on the same page, which advances the discussion rather than correcting or stagnating on the same points. Several participants also mentioned how the system may be used to ask for and exchange information (P4, P10, P14, P17), which pushes the conversation to be constructive as well as argumentative. Though the system may be centered on the called-out individual, many participants felt that the presence of the system's features could also lead to reduced harassing behavior in the long term.

Social computing research has also explored the possibility of facilitating healthy discussions by gaining a better understanding of others and reducing hostility. Systems such as ConsiderIt [54] and Reflect [55] experimented with 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 [73]. Re:SPect builds upon such approaches by promoting balanced opinion perception and reduced hostility in online environments. We also argue that features that enrich the context of a discussion could help provide additional nuances that could not be represented in previous social media-based interactions. Kim et al. have previously explored the possibility of priming users to contextual information about a post [51] to encourage open discussion and reduce animosity. Similarly, we encourage future work to further explore features flexibly providing nuanced context to a discussion so as to facilitate healthy, open discussions online.



		Level of Initiative	
		Passive	**Active**
Engagement Strategy	**Direct**	● Asking for support from peers	● Responding individually ● Taking legal action
	Indirect	● No response ● Deleting the post or account ● Turning into a private account	● Responding through a public post ● Posting a public apology/amendment

**Categories are not mutually exclusive*

Fig. 6. Classification of Responses to Online Harassment

7.2 Responding to Networked Online Harassment

Our observations yielded insights into the methods Twitter users use to respond to networked online harassment. Previous work on online harassment has explored the various motivations for protective strategies and responses used by victims of harassment, as well as how these strategies are evaluated by the user [63]. Depending on the experiences, preferences, or situational variables, users may employ different response strategies, which can lead to a diverse array of consequences. Thus, we argue that classifying the responses to online harassment could be beneficial in validating the victims' experiences [9] as well as understanding the motivations behind them. In this section, we discuss the key factors in effectively responding to online harassment, as well as how users respond to networked harassment.

There are a variety of methods that users can take to protect themselves from harassment, such as reporting [60, 63, 68] or blocking users [14, 46], deleting the post, changing the privacy settings of or deleting one's account, and taking legal action [52], among others. In the context of networked harassment and public calling out, another important factor is to clarify misunderstandings and share missing or additional context (Section 6.3.1), or stopping the spread of posts (Section 6.1.1) to prevent further harassment. Kim et al. have classified the response tactics of called-out individuals into active and passive responses where the individual either acknowledges and engages with the criticizing content or chooses to avoid it [52].

We expand upon this framework to propose another variable for classifying responses to online harassment: the engagement level of the responses. We build upon the passive-active dichotomy described earlier by Kim et al., which specifies how much initiative does the individual take to respond to the harassment. We consider this in combination with the engagement level variable: Do they reply directly to the perpetrators, or address them indirectly through posts targeting the general public? Our suggested model and examples of harassment responses as reported by the participants of our study are depicted in Figure 6.

7.2.1 Re:SPect Promoting More Active Responses. While the decision of which response measure to take varies, and will largely depend on individual differences, there were some patterns examined in the interviews that could suggest future design implications for anti-harassment systems. For example, after using Re:SPect, many participants reported that they would take a more active approach to handling networked harassment. P4 said that their preferred method of responding to networked harassment was "*just to delete it, or turn off notifications, and wait until people don't care anymore*." However, after being introduced to Re:SPect, they said that the array of choices that

Re:SPect provided them would encourage them to try out more active response tactics such as adding context or specifying visibility boundaries. We note that in our participants, passive responses were usually taken out of a sense of helplessness surrounding networked online harassment. Thus, future work could use the suggested model as a frame of reference and aim to provide methods that allow victims of harassment to take up more active response measures than passive ones.

7.2.2 Allowing for Indirect Responses to Networked Harassment. While our participants did not show a clear preference towards either direct or indirect response measures, Re:SPect did allow for preferences toward either direction. As Re:SPect was designed to provide an additional layer of protection for the users, our design mostly focused on promoting more indirect responses. This could have the benefit of suppressing workarounds perpetrators may take to determinedly harass the victim, as is known to happen in several networked harassment situations [52, 83]. However, some participants did note that they would prefer a more direct method of communicating with the harassers. For example, while we introduced the post flag feature to promote an indirect method of context sharing, some participants mentioned that they wanted to send notifications to their harassers regarding the flags. They noted that this was so that they would understand that they had contributed to harassment and reconsider their actions.

We can explore this by focusing on the aspect of reputation in social media. As noted before, a major issue of networked harassment is that it is often difficult or impossible to control how far it spreads [66], often beyond the direct comments made by perpetrators [52]. Indirect responses, then, could function as a method to respond and express oneself to the larger, potential audience. This, in connection to the three types of stakeholders in online harassment – victim, perpetrator, bystander [24, 25, 98] – direct responses would be when the victim engages with the perpetrator and indirect responses when they engage with the bystanders.

Our results imply that while these motivations are not mutually exclusive, the driving factors and response patterns may differ significantly. In exploring user-centric responses to harassment, understanding the victim's motivations and demands towards these third parties could be beneficial to better effectively leverage community support in response to harassment. Examples of this could include requesting moderation help from close peers [64], encouraging bystanders to step in [24, 98], or simply preventing misunderstandings. If we obtain a better understanding of these motivations and the potential impact surrounding bystander audiences, it could enable more effective and flexible methods of responding at scale.

7.3 Increasing Perceived Agency and Self-Efficacy

The findings from our user study indicate that Re:SPect's features can be effective in solving the problems related to online harassment; however, this is not because the features suggested will eradicate harassment. While a large portion of online antisocial behavior is driven by situational variables [21], people who are determined to harass, either due to malicious intent or simply to cause chaos [53, 77] cannot be stopped completely. In fact, several participants in our study noted that anti-harassment systems and features, no matter how well-executed, will still not be able to remove all harassment from the platform as people will eventually find a way to circumvent them.

Even so, participants felt empowered and safer as a result of using Re:SPect. Part of it could be attributed to the fact that Re:SPect's features could make harassment a much more high-maintenance task for the perpetrators. The existence of protection and response measures cut back on the possibility that people may participate in harassment unintentionally [46, 52] or by conforming to their peers [14]. More importantly, participants also noted that the existence of Re:SPect, as well as the breadth of online harassment scenarios that it can cover, increased the *perceived agency* of the participants. Perceived response efficacy and self-efficacy provided by the existence and knowledge of effective response measures can benefit victims of harassment by encouraging them to take action [63]. Social media users often censor their postings in fear of being targeted for online harassment, even when they had not previously been harassed [60]. The reassurance that one will be able to respond to harassment even if it happens can make them feel safer, freer, and less anxious about potential negative responses to their posts.

7.4 The Future of Networked Harassment Moderation in the Decentralized Web

While this study focused on replicating and improving harassment-related experiences on Twitter, the implications of our study can extend beyond the platform. Particularly, we focus on the relevance of the current work in decentralized social media and moderation practices, also known as the fediverse. Following the acquisition of the microblogging platform by Elon Musk in October 2022, there have been significant changes in platform policy, management, and overall user experience [4, 78, 97, 102]. In response to this, a movement to migrate to alternative microblogging platforms has emerged [44, 106]. While there are certainly examples of alternative platforms that also follow a centralized protocol, Meta's Threads[2] being the most recent example, it is interesting to note that many of the prominent alternatives are part of the decentralized web, such as Mastodon[3], Bluesky[4], and more. As our design also focused on a more individualized, distributed form of harassment, the design implications from Re:SPect could also be applied to such decentralized social media platforms.

Moderation in the fediverse is unique in that the policies differ according to each instance, or the specific network of users involved [2, 84]. This has the effect of allowing for more flexible moderation decisions that better fit the characteristics and expectations of each community [7, 84], as opposed to the centralized moderation protocol used by private platforms. Users are also given the choice to decide which set of moderation policies they follow by being able to freely join and leave instances [28, 84]. However, the burden of moderation falling solely on the instances' moderators mean that it presents a larger burden to the individuals involved, especially considering that it is often conducted through volunteer labor [2, 28]. Harassment may also persist across instances, for example, exploiting instances that are not heavily moderated or even forming instances that are dedicated to harassment and toxic behavior [13, 18, 100]. Thus, we emphasize the importance of allowing individual users to control and moderate their own audiences, as we have demonstrated through Re:SPect. The fediverse may take this into account so as to reduce the burden of volunteer moderators and empower individual users to respond against networked harassment.

7.5 Limitations and Future Work

While we evaluate the feasibility and users' perceptions of Re:SPect through a simulated user study, we were not able to fully implement and deploy an end-to-end version that could be applied to 'real' harassment situations. Real-time factors such as how fast are the responses coming are often taken into consideration when assessing the severity of harassment but were not implemented in our study design. We have attempted to alleviate these concerns by providing both preemptive and reactive scenarios of networked harassment, but as they were also both based on a single moment in time, they may not fully represent what victims of networked harassment are going through. We also note that introducing different and more diverse sample situations in the scenario-based study could have prompted more diverse potential use cases and implications for Re:SPect. Future work

[2] https://www.threads.net/
[3] https://joinmastodon.org/
[4] https://blueskyweb.org/

could improve upon the generalizability of our findings by introducing more complex situations, such as different catalysts of harassment, complexity, and scale of the network involved.

In addition, as our evaluation of Re:SPect focused more on the proof-of-concept of anti-harassment social media design, some aspects specific to implementation were not applied. For example, with regards to **D2** (Provide a succinct, digestible summary of user comments), we advised the participants to assume that Re:SPect would auto-generate clusters of opinions. In this case, performance metrics of the model, such as accuracy, could impact the feasibility and trustworthiness of the feature. Future work could explore how different clustering models or varying levels of performance could impact user perceptions and usage patterns.

Finally, it was brought to our attention that some of our features, such as users limiting the visibility boundaries of their posts, may go against the interests of social media platform companies. It is true that many commercial platforms such as Twitter, Facebook, Instagram, and YouTube focus on engagement and amplification of content. However, unsavory experiences surrounding online harassment are often a reason for reduced engagement or leaving the platform entirely [74], deteriorating the quality of discourse. Thus, it would be in the interest of such companies to apply individualized anti-harassment designs as suggested in this paper. While this paper focuses primarily on Twitter, it may be interesting to see how the insights from this paper could be applied to other platforms.

8 CONCLUSION

In this paper, we explored the design and development of anti-harassment tools on social media, specifically focusing on networked harassment. Through three sessions of design workshops, we revealed key elements and design goals to consider when designing to prevent networked harassment on Twitter. These included the accurate representation of the poster's intent and content, ways to stop the amplification of posts, as well as providing granularity and specificity in the post settings. We designed Re:SPect, a system promoting scalable and active responses from victims of networked harassment, and evaluated it with 18 participants through a speculative scenario-based study. Our findings suggest that providing scalable response measures such as context nudges and mass-response features was effective in reducing anxiety and the feeling of helplessness in networked harassment. Building upon these insights, we introduce implications and theoretical frameworks upon which we could develop more effective solutions to networked harassment.

REFERENCES

[1] Renee Nicole Allen. 2021. From academic freedom to cancel culture: Silencing black women in the legal academy. *UCLA L. Rev.* 68 (2021), 364.

[2] Ishaku Hassan Anaobi, Aravindh Raman, Ignacio Castro, Haris Bin Zia, Damilola Ibosiola, and Gareth Tyson. 2023. Will Admins Cope? Decentralized Moderation in the Fediverse. In *Proceedings of the ACM Web Conference 2023*. 3109–3120.

[3] Shaowen Bardzell and Jeffrey Bardzell. 2011. Towards a feminist HCI methodology: social science, feminism, and HCI. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 675–684.

[4] Jenae Barnes. 2023. Twitter Ends Its Free API: Here's Who Will Be Affected. https://www.forbes.com/sites/jenaebarnes/2023/02/03/twitter-ends-its-free-api-heres-who-will-be-affected/ Section: Business.

[5] Michael S Bernstein, Eytan Bakshy, Moira Burke, and Brian Karrer. 2013. Quantifying the invisible audience in social networks. In *Proceedings of the SIGCHI conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 21–30.

[6] Aparajita Bhandari, Marie Ozanne, Natalya N Bazarova, and Dominic DiFranzo. 2021. Do You Care Who Flagged This Post? Effects of Moderator Visibility on Bystander Behavior. *Journal of Computer-Mediated Communication* 26, 5 (2021), 284–300.

[7] Haris Bin Zia, Aravindh Raman, Ignacio Castro, Ishaku Hassan Anaobi, Emiliano De Cristofaro, Nishanth Sastry, and Gareth Tyson. 2022. Toxicity in the decentralized web and the potential for model sharing. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 6, 2 (2022), 1–25.

[8] Lindsay Blackwell, Tianying Chen, Sarita Schoenebeck, and Cliff Lampe. 2018. When Online Harassment Is Perceived as Justified. *Proceedings of the International AAAI Conference on Web and Social Media* 12, 1 (2018), 10.

[9] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW (Dec. 2017), 1–19. https://doi.org/10.1145/3134659

[10] Yarimar Bonilla and Jonathan Rosa. 2015. #Ferguson: Digital protest, hashtag ethnography, and the racial politics of social media in the United States. *American ethnologist* 42, 1 (2015), 4–17.

[11] Gwen Bouvier. 2020. Racist call-outs and cancel culture on Twitter: The limitations of the platform's ability to define issues of social justice. *Discourse, Context & Media* 38 (Dec. 2020), 100431. https://doi.org/10.1016/j.dcm.2020.100431

[12] danah boyd. 2008. Why youth (heart) social network sites: The role of networked publics in teenage social life. *YOUTH, IDENTITY, AND DIGITAL MEDIA, David Buckingham, ed., The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning, The MIT Press, Cambridge, MA* 2007-16, 1 (2008), 119–142.

[13] Joshua Braun. 2023. Journalism, Media Research, and Mastodon: Notes on the Future. *Digital Journalism* (2023), 1–8.

[14] André Brock Jr. 2020. *Distributed Blackness: African American Cybercultures.* NYU Press.

[15] Nicholas Brody. 2021. Bystander Intervention in Cyberbullying and Online Harassment: The Role of Expectancy Violations. *International Journal of Communication* 15, 0 (Jan. 2021), 21. https://ijoc.org/index.php/ijoc/article/view/14169 Number: 0.

[16] Nicholas Brody and Anita L Vangelisti. 2016. Bystander intervention in cyberbullying. *Communication Monographs* 83, 1 (2016), 94–119.

[17] Amanda Burgess-Proctor, Justin W Patchin, and Sameer Hinduja. 2009. Cyberbullying and online harassment: Reconceptualizing the victimization of adolescent girls. *Female crime victims: Reality reconsidered* (2009), 153–175.

[18] Derek Caelin. 2022. Decentralized networks vs the trolls. In *Fundamental challenges to global peace and security: The future of humanity*. Springer, 143–168.

[19] Jie Cai and Donghee Yvette Wohn. 2019. What are Effective Strategies of Handling Harassment on Twitch? Users' Perspectives. In *Conference companion publication of the 2019 on computer supported cooperative work and social computing*. 166–170.

[20] Kalyani Chadha, Linda Steiner, Jessica Vitak, and Zahra Ashktorab. 2020. Women's responses to online harassment. *International journal of communication* 14 (2020), 19.

[21] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 1217–1230. https://doi.org/10.1145/2998181.2998213

[22] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. Antisocial behavior in online discussion communities. In *Proceedings of the international aaai conference on web and social media*, Vol. 9. 61–70.

[23] Meredith D. Clark. 2020. DRAG THEM: A brief etymology of so-called "cancel culture". *Communication and the Public* 5, 3-4 (Sept. 2020), 88–92. https://doi.org/10.1177/2057047320961562

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

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

[26] Maeve Duggan. 2017. Online harassment 2017. (2017).

[27] Satu Elo and Helvi Kyngäs. 2008. The qualitative content analysis process. *Journal of advanced nursing* 62, 1 (2008), 107–115.

[28] Ksenia Ermoshina and Francesca Musiani. 2022. Safer spaces by design? Federated architectures and alternative socio-technical models for content moderation. In *Annual Symposium of the Global Internet Governance Academic Network (GigaNet)*.

[29] Michelle Ferrier and Nisha Garud-Patkar. 2018. TrollBusters: Fighting online harassment of women journalists. *Mediating misogyny: Gender, technology, and harassment* (2018), 311–332.

[30] Casey Fiesler, Shannon Morrison, and Amy S Bruckman. 2016. An archive of their own: A case study of feminist HCI and values in design. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2574–2585.

[31] Jerry Finn. 2004. A survey of online harassment at a university campus. *Journal of Interpersonal violence* 19, 4 (2004), 468–483.

[32] Johnathan Flowers. 2022. The Whiteness of Mastodon. https://techpolicy.press/the-whiteness-of-mastodon/

[33] Jesse Fox and Wai Yen Tang. 2017. Women's experiences with general and sexual harassment in online video games: Rumination, organizational responsiveness, withdrawal, and coping strategies. *New media & society* 19, 8 (2017), 1290–1307.

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

[35] R Stuart Geiger. 2016. Bot-based collective blocklists in Twitter: the counterpublic moderation of harassment in a networked public space. *Information, Communication & Society* 19, 6 (2016), 787–803.

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

[37] Tarleton Gillespie. 2018. *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*.

[38] 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. 2017. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on web science conference*. 229–233.

[39] Chandell Enid Gosse and Victoria Jane O'Meara. 2018. Blockbotting dissent": Publics, counterpublics, and algorithmic public sphere (s). *Stream: Inspiring Critical Thought* 10, 1 (2018), 3–11.

[40] Joshua Guberman, Carol Schmitz, and Libby Hemphill. 2016. Quantifying toxicity and verbal violence on Twitter. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*. 277–280.

[41] Oliver L Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 2021. Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–35.

[42] Yim Hyun-su. 2018. #MeToo, feminism dominated Twitter in 2018. https://www.koreaherald.com/view.php?ud=20181206000775

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

[44] Ujun Jeong, Paras Sheth, Anique Tahir, Faisal Alatawi, H Russell Bernard, and Huan Liu. 2023. Exploring Platform Migration Patterns between Twitter and Mastodon: A User Behavior Study. *arXiv preprint arXiv:2305.09196* (2023).

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

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

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

[48] Franccesca Kazerooni, Samuel Hardman Taylor, Natalya N Bazarova, and Janis Whitlock. 2018. Cyberbullying bystander intervention: The number of offenders and retweeting predict likelihood of helping a cyberbullying victim. *Journal of Computer-Mediated Communication* 23, 3 (2018), 146–162.

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

[50] Do Own Donna Kim, Nathaniel Ming Curran, and Hyun Tae Calvin Kim. 2020. Digital Feminism and Affective Splintering: South Korean Twitter Discourse on 500 Yemeni Refugees. *International Journal of Communication* 14 (2020), 19.

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

[52] Haesoo Kim, HaeEun Kim, Juho Kim, and Jeong-woo Jang. 2022. When Does it Become Harassment? An Investigation of Online Criticism and Calling Out in Twitter. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–32.

[53] Ben Kirman, Conor Lineham, and Shaun Lawson. 2012. Exploring mischief and mayhem in social computing or: how we learned to stop worrying and love the trolls. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*. 121–130.

[54] Travis Kriplean, Jonathan Morgan, Deen Freelon, Alan Borning, and Lance Bennett. 2012. Supporting reflective public thought with considerit. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. Association for Computing Machinery, New York, NY, USA, 265–274.

[55] Travis Kriplean, Michael Toomim, Jonathan Morgan, Alan Borning, and Amy Ko. 2012. Is this what you meant? Promoting listening on the web with reflect. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1559–1568.

[56] Colette Langos. 2012. Cyberbullying: The Challenge to Define. *Cyberpsychology, Behavior, and Social Networking* 15, 6 (June 2012), 285–289. https://doi.org/10.1089/cyber.2011.0588

[57] Ho Lee, Jaewon Choi, Kyung Kyu Kim, and Ae Ri Lee. 2014. Impact of anonymity on information sharing through internal psychological processes: A case of South Korean online communities. *Journal of Global Information Management (JGIM)* 22, 3 (2014), 57–77.

[58] Song Mi Lee, Andrea K Thomer, and Cliff Lampe. 2022. The Use of Negative Interface Cues to Change Perceptions of Online Retributive Harassment. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–23.

[59] Una Lee and Dann Toliver. 2017. Building Consentful Tech. 2017. http://www.consentfultech.io/wp-content/uploads/2019/10/Building-Consentful-Tech.pdf

[60] Amanda Lenhart, Michele Ybarra, Kathryn Zickuhr, and Myeshia Price-Feeney. 2016. *Online harassment, digital abuse, and cyberstalking in America*. Data and Society Research Institute.

[61] Rebecca Lewis, Alice E Marwick, and William Clyde Partin. 2021. "We Dissect Stupidity and Respond to It": Response Videos and Networked Harassment on YouTube. *American Behavioral Scientist* 65, 5 (2021), 735–756.

[62] Paul Benjamin Lowry, Jun Zhang, Chuang Wang, and Mikko Siponen. 2016. Why do adults engage in cyberbullying on social media? An integration of online disinhibition and deindividuation effects with the social structure and social learning model. *Information Systems Research* 27, 4 (2016), 962–986.

[63] May O Lwin, Benjamin Li, and Rebecca P Ang. 2012. Stop bugging me: An examination of adolescents' protection behavior against online harassment. *Journal of adolescence* 35, 1 (2012), 31–41.

[64] Kaitlin Mahar, Amy X. Zhang, and David Karger. 2018. Squadbox: A Tool to Combat Email Harassment Using Friendsourced Moderation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–13. https://doi.org/10.1145/3173574.3174160

[65] Thabo Mahlangu, Chunling Tu, and Pius Owolawi. 2018. A review of automated detection methods for cyberbullying. In *2018 International Conference on Intelligent and Innovative Computing Applications (ICONIC)*. IEEE, 1–5.

[66] Alice E Marwick. 2021. Morally motivated networked harassment as normative reinforcement. *Social Media+ Society* 7, 2 (2021), 20563051211021378.

[67] Alice E Marwick and Robyn Caplan. 2018. Drinking male tears: Language, the manosphere, and networked harassment. *Feminist Media Studies* 18, 4 (2018), 543–559.

[68] J Nathan Matias, Amy Johnson, Whitney Erin Boesel, Brian Keegan, Jaclyn Friedman, and Charlie DeTar. 2015. Reporting, reviewing, and responding to harassment on Twitter. *arXiv preprint arXiv:1505.03359* (2015), 63 pages.

[69] Colten Meisner. 2023. Networked Responses to Networked Harassment? Creators' Coordinated Management of "Hate Raids" on Twitch. *Social Media+ Society* 9, 2 (2023), 20563051231179696.

[70] Ersilia Menesini and Annalaura Nocentini. 2009. Cyberbullying Definition and Measurement: Some Critical Considerations. *Zeitschrift für Psychologie / Journal of Psychology* 217, 4 (Jan. 2009), 230–232. https://doi.org/10.1027/0044-3409.217.4.230

[71] Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2019. Tackling online abuse: A survey of automated abuse detection methods. *arXiv preprint arXiv:1908.06024* (2019).

[72] Lisa Nakamura. 2015. The unwanted labour of social media: Women of colour call out culture as venture community management. *New Formations* 86, 86 (2015), 106–112.

[73] Matti Nelimarkka, Jean Philippe Rancy, Jennifer Grygiel, and Bryan Semaan. 2019. (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 (2019), 1–25.

[74] Fayika Farhat Nova, Michael Ann DeVito, Pratyasha Saha, Kazi Shohanur Rashid, Shashwata Roy Turzo, Sadia Afrin, and Shion Guha. 2021. " Facebook Promotes More Harassment" Social Media Ecosystem, Skill and Marginalized Hijra Identity in Bangladesh. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–35.

[75] Jessica A Pater, Moon K Kim, Elizabeth D Mynatt, and Casey Fiesler. 2016. Characterizations of online harassment: Comparing policies across social media platforms. In *Proceedings of the 19th international conference on supporting group work*. 369–374.

[76] Sai Teja Peddinti, Keith W Ross, and Justin Cappos. 2014. " On the internet, nobody knows you're a dog" a twitter case study of anonymity in social networks. In *Proceedings of the second ACM conference on Online social networks*. 83–94.

[77] Whitney Phillips. 2015. *This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture.* Mit Press.

[78] The Associated Press. 2022. Musk's Twitter has dissolved its Trust and Safety Council. *NPR* (Dec. 2022). https://www.npr.org/2022/12/12/1142399312/twitter-trust-and-safety-council-elon-musk

[79] Kususanto Prihadi, Yen Ling Hui, Melissa Chua, and Calvin KW Chang. 2019. Cyber-Victimization and Perceived Depression: Serial Mediation of Self-Esteem and Learned-Helplessness. *International Journal of Evaluation and Research in Education* 8, 4 (2019), 563–574.

[80] Martin J Riedl, Katie Joseff, Stu Soorholtz, and Samuel Woolley. 2022. Platformed antisemitism on Twitter: Anti-Jewish rhetoric in political discourse surrounding the 2018 US midterm election. *new media & society* (2022), 14614448221082122.

[81] Sarah Roberts. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media.* Yale University Press.

[82] Sarah T Roberts. 2016. Commercial content moderation: Digital laborers' dirty work. (2016).

[83] Jon Ronson. 2016. *So You've Been Publicly Shamed.* Riverhead Books.

[84] Alan Z Rozenshtein. 2022. Moderating the Fediverse: Content Moderation on Distributed Social Media. (2022).

[85] Niloufar Salehi, Roya Pakzad, Nazita Lajevardi, and Mariam Asad. 2023. Sustained Harm Over Time and Space Limits the External Function of Online Counterpublics for American Muslims. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–24.

[86] Michael Salter. 2013. Justice and revenge in online counter-publics: Emerging responses to sexual violence in the age of social media. *Crime, Media, Culture* 9, 3 (2013), 225–242.

[87] Amit M Schejter and Noam Tirosh. 2015. "Seek the meek, seek the just": Social media and social justice. *Telecommunications policy* 39, 9 (2015), 796–803.

[88] Sarita Schoenebeck, Oliver L Haimson, and Lisa Nakamura. 2021. Drawing from justice theories to support targets of online harassment. *New Media & Society* 23, 5 (May 2021), 1278–1300. https://doi.org/10.1177/1461444820913122

[89] Sarita Schoenebeck, Cliff Lampe, and Penny Triệu. 2023. Online Harassment: Assessing Harms and Remedies. *Social Media+ Society* 9, 1 (2023), 20563051231157297.

[90] Sarita Schoenebeck, Carol F. Scott, Emma Grace Hurley, Tammy Chang, and Ellen Selkie. 2021. Youth Trust in Social Media Companies and Expectations of Justice: Accountability and Repair After Online Harassment. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1 (April 2021), 1–18. https://doi.org/10.1145/3449076

[91] Joseph Seering. 2020. Reconsidering self-moderation: the role of research in supporting community-based models for online content moderation. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–28.

[92] Joseph Seering, Tianmi Fang, Luca Damasco, Mianhong'Cherie' Chen, Likang Sun, and Geoff Kaufman. 2019. 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.* 1–14.

[93] Martin EP Seligman. 1972. Learned helplessness. *Annual review of medicine* 23, 1 (1972), 407–412.

[94] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J Riedl, and Matthew Lease. 2021. The psychological well-being of content moderators: the emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the 2021 CHI conference on human factors in computing systems.* 1–14.

[95] Leo Graiden Stewart, Ahmer Arif, A Conrad Nied, Emma S Spiro, and Kate Starbird. 2017. Drawing the lines of contention: Networked frame contests within# BlackLivesMatter discourse. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–23.

[96] Sharifa Sultana, Mitrasree Deb, Ananya Bhattacharjee, Shaid Hasan, SM Raihanul Alam, Trishna Chakraborty, Prianka Roy, Samira Fairuz Ahmed, Aparna Moitra, M Ashraful Amin, et al. 2021. 'Unmochon': A Tool to Combat Online Sexual Harassment over Facebook Messenger. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* 1–18.

[97] Josh Taylor and Dan Milmo. 2023. How Twitter's new drastic changes will affect what users can view on the site. *The Guardian* (July 2023). https://www.theguardian.com/technology/2023/jul/03/how-twitter-new-changes-will-affect-users-rate-limited-limit-exceeded-restrictions

[98] Samuel Hardman Taylor, Dominic DiFranzo, Yoon Hyung Choi, Shruti Sannon, and Natalya N Bazarova. 2019. Accountability and empathy by design: Encouraging bystander intervention to cyberbullying on social media. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–26.

[99] Dias Oliva Thiago, Antonialli Dennys Marcelo, and Alessandra Gomes. 2021. Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online. *Sexuality & culture* 25, 2 (2021), 700–732.

[100] Kevin Veale and Kevin Veale. 2020. Problematic Tools and Platform Complicity. *Gaming the Dynamics of Online Harassment* (2020), 107–128.

[101] Jessica Vitak, Kalyani Chadha, Linda Steiner, and Zahra Ashktorab. 2017. Identifying Women's Experiences With and Strategies for Mitigating Negative Effects of Online Harassment. In *Proceedings of the 2017 ACM Conference*

on Computer Supported Cooperative Work and Social Computing. ACM, Portland Oregon USA, 1231–1245. https://doi.org/10.1145/2998181.2998337

[102] Kurt Wagner. 2023. Twitter cuts workers addressing hate speech and trust and safety as Elon Musk's chaotic revamp continues. https://fortune.com/2023/01/07/twitter-cuts-workers-hate-speech-trust-safety-elon-musk-revamp/

[103] Magdalena Wojcieszak, Andreu Casas, Xudong Yu, Jonathan Nagler, and Joshua A Tucker. 2021. Echo chambers revisited: The (overwhelming) sharing of in-group politicians, pundits and media on Twitter. (2021).

[104] Randy Yee Man Wong, Christy MK Cheung, Bo Xiao, and Jason Bennett Thatcher. 2021. Standing up or standing by: Understanding bystanders' proactive reporting responses to social media harassment. *Information Systems Research* 32, 2 (2021), 561–581.

[105] Austin P Wright, Omar Shaikh, Haekyu Park, Will Epperson, Muhammed Ahmed, Stephane Pinel, Duen Horng Chau, and Diyi Yang. 2021. RECAST: Enabling User Recourse and Interpretability of Toxicity Detection Models with Interactive Visualization. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–26.

[106] Haris Bin Zia, Jiahui He, Aravindh Raman, Ignacio Castro, Nishanth Sastry, and Gareth Tyson. 2023. Flocking to mastodon: Tracking the great twitter migration. *arXiv preprint arXiv:2302.14294* (2023).