International
Communication
Association

OXFORD

To intervene or not to intervene: young adults' views on when and how to intervene in online harassment

Anna Davidovic[1,*], Catherine Talbot [2], Catherine Hamilton-Giachritsis[3], Adam Joinson[1]

[1]School of Management, University of Bath, Bath and North East Somerset, Bath, UK
[2]School of Psychology, University of Bournemouth, Poole, Dorset, UK
[3]Department of Psychology, University of Bath, UK, Bath and North East Somerset, Bath, UK
*Corresponding author: Anna Davidovic. Email: ad2150@bath.ac.uk

Abstract

Incidents of online harassment are increasing and can have significant consequences for victims. Witnesses ("digital bystanders") can be crucial in identifying and challenging harassment. This study considered when and how young adults intervene online, with the aim of understanding the applicability of existing theoretical models (i.e., Bystander Intervention Model; Response Decision-Making Framework). Thematic analysis of eight focus groups (UK community sample, $N = 67$, 18–25 years) resulted in five themes: Noticing and Interpreting the Harassment, Perceived Responsibility for Helping, Consequences of Intervening, Perceived Ability to Make a Difference, and Deciding How to Help. The online context amplified offline preferences, such as greater preference for anonymity and perceived costs of intervention (e.g., social costs). Intervention strategies varied in visibility and effort, preferring "indirect" micro-interventions focused on supporting victims. A new, merged model specific to digital bystanders is proposed, with implications for the design and messaging on social networking sites discussed.

Lay Summary

What influences someone to step in when they see harassment taking place online? We asked 67 young adults about their experiences of online harassment using focus groups and identified key themes. We found that bystanders are generally reluctant to intervene unless the harassment is obvious, severe, and when they knew the victim well. When they did intervene, participants preferred to step in discreetly through reporting (to social networking sites) or private messaging to offer emotional support. Reporting was not always seen in a positive light, with participants describing it as an "empty" experience due to the lack of feedback about what happened next. Bystanders were concerned about making the situation worse, becoming a victim themselves, and not having the right skills to intervene. In the case of celebrities and influencers, online harassment was seen as "part of the job." This was amplified by a feeling of detachment due to the fact they cannot physically see the victim online. Our study highlights several barriers that discourage bystanders from intervening and suggests ways of removing these barriers through design and messaging. We propose a new way of conceptualizing online interventions along a "spectrum" varying in levels of visibility and effort.

Keywords: social networking sites, online communities, qualitative methods, young adults, bystanders

Abusive online behavior, such as trolling and flaming, has been the subject of research for the past 40 years in the computer-mediated communication (CMC) field (Kiesler et al., 1984). While online harassment is a much-disputed term (Marwick, 2021), we define it as targeted abuse or harmful behavior directed at another individual user through CMC. Online harassment is often viewed as a spectrum of behaviors varying in severity (e.g., purposeful embarrassment, stalking, physical threats, sexual harassment; see Pew Research Center, 2021). Online harassment is now increasingly widespread and severe, with 41% of surveyed American adults experiencing some form of harassment (Pew Research Center, 2021). Furthermore, compared against the 2017 survey, harassment has become more severe in nature (e.g., physical threats, stalking, and sustained harassment) (Pew Research Center, 2021; UK Council for Internet Safety, 2019). The public and "permanent" nature of online abuse can lead to long term, and potentially devastating, consequences for victims (Dillon & Bushman, 2015). For example, victims of cyberbullying report higher levels of depressive symptoms (Perren et al., 2010) and intrusive thoughts (Liu et al., 2020) compared to victims of traditional in-person bullying.

Detecting and tackling online harassment—why bystanders?

Alongside increasing interest in both the prevalence and detection of online harms (Rosa et al., 2019), increasing attention has been focused on the role of "digital bystanders"—i.e., those who witness online abuse or harmful communications, and have the opportunity to intervene (e.g., Difranzo et al., 2018; Dillon & Bushman, 2015; Obermaier et al., 2016). Digital bystanders (also referred to as "cyber-bystanders" in the cyberbullying literature) have a potentially crucial role in challenging and reporting online abuse and harassment, particularly in light of reported challenges in the automated, real-time detection of problematic content (Rosa et al., 2019). Furthermore, witnessing another intervene in public (e.g., posting a direct challenge to the abuse) can role model different intervention strategies, thereby encouraging others to intervene (Anderson et al., 2014; Marwick, 2021).

Social networking sites (SNS) offer opportunities for bystanders to intervene anonymously via reporting systems and the capability to document an incident readily and easily (e.g., taking screenshots). For these reasons, we might expect that bystanders would be motivated to intervene. Yet, survey research suggests that this is not the case, with the majority of

those questioned (70%) reporting that they opt not to intervene (Pew Research Center, 2017). During the last decade, experimental research has also found consistently low levels of intervention in online settings (Bhandari et al., 2021; Difranzo et al., 2018; Dillon & Bushman, 2015), particularly in relation to "public" interventions. The relatively low rate of bystander intervention is a phenomenon previously identified in *offline* settings (Latané and Darley, 1970) and was the focus of considerable early research in the 1960s that culminated in a "Bystander Intervention Model" (Latané and Darley, 1970), which sought to identify the potential reasons for non-intervention.

The bystander intervention model

Bystanders have been traditionally studied in relation to in-person emergencies, such as physical assault (Allison & Bussey, 2016). Empirical work in this field has largely centered around the well-established bystander intervention model (BIM) (Latané and Darley, 1970) consisting of five underpinning stages (see Figure 1).

According to the model, each stage is associated with situational factors that can encourage or deter bystander intervention (e.g., lack of skills or knowledge, "evaluation apprehension"). In their meta-analytic review ($k = 7,700$), Fischer et al. (2011) synthesized this extensive field of research and found that bystanders are more likely to help in dangerous emergencies, when perpetrators are present, and when the "costs" to themselves are nonphysical. More recently, Robinson et al. (2022) conducted a qualitative interpretive metasynthesis of bystander intervention in the sexual violence literature, identifying five key themes: the impact of alcohol (acting as both an enabler of or deterrent to intervention depending on the context), peer perceptions (e.g., rape myths acceptance as a key barrier), beliefs around responsibility (acting as both an enabler or deterrent depending on perceptions of moral duty), situational factors (e.g., the role of
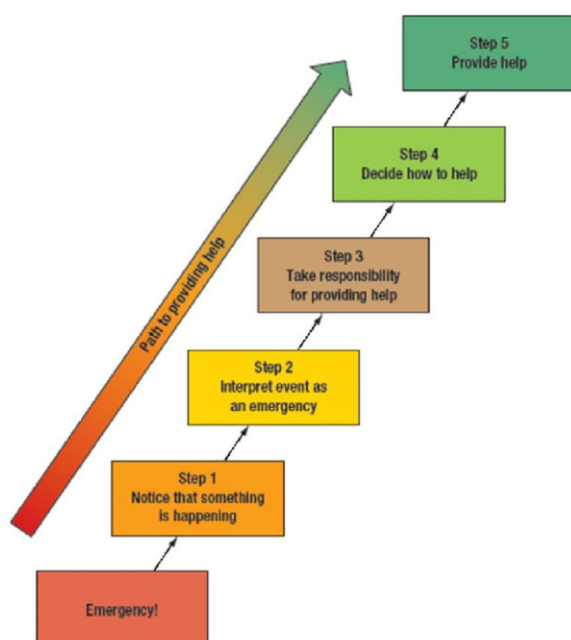
Figure 1. Bystander intervention model (adapted from Latané & Darley, 1970, as cited in Dillon & Bushman, 2015).

body language), and the role of friendship (as a key enabler of helping).

Barriers and enablers of bystander intervention

In the last decade, research has moved toward understanding ways in which bystander intervention can be enabled or supported—both offline and online. For instance, Fenton et al. (2016) reviewed 67 studies of bystander intervention programs and concluded that—alongside evidence that such programs are effective in reducing offline abuse and harassment—active bystander intervention is associated with increases in victim-empathy, confidence (and self-efficacy), and a willingness to accept personal responsibility to act. However, as is apparent in the Robinson et al. (2022) qualitative metasynthesis, some features can act as either an enabler or a barrier, depending on the context. For example, alcohol can be both a facilitator or an inhibitor of bystander intervention. Alcohol can reduce inhibitions and increase confidence of bystanders (facilitating bystander intervention), or it can act as a barrier by increasing ambiguity and impairing decision making (reducing bystander intervention).

Studies in digital environments tend to replicate these same effects. For instance, and in keeping with the predictions of the original BIM, increased personal responsibility, incident severity, and number of offenders, have all been found to increase the likelihood of a victim receiving help in a digital setting (Butler et al., 2022; Kazerooni et al., 2018; Rudnicki et al., 2022; Wang, 2021). Notably, in their scenario-based study, Bastiaensens et al. (2014) found that incident severity and a prior relationship with the victim increased secondary school students' ($N = 453$) behavioral intentions to help. Similarly, Brody (2013) reported that closeness to the victim and the anonymity of the bystander increased intervention in both vignette-based and experimental studies.

Replicating previous work in offline settings, empathy with the victim has consistently been found to predict online intervention (e.g., Van Cleemput et al., 2014), although there may be differences in terms of cognitive vs. affective empathy (Barlińska et al., 2018). Wang (2021) found that those higher in dispositional empathy were more likely to intervene (both publicly and privately) in a scenario based cyber-bullying experiment, while Rudnicki et al. (2020) found that giving participants oxytocin to increase empathy led to a decrease in their reported acceptance of "celebrity bashing" in online settings. However, in a 3-day simulation study, Taylor et al. (2019) found no evidence that higher empathic concern led to more "flagging" (one of the intervention options within the study). Interestingly, seeing a "feeling prompt" (e.g., "how is X feeling?") predicted greater empathic concern, which in turn, did lead to higher intention to "like" the victims' posts, suggesting that different types of intervention may be motivated by different factors and perhaps the diverse findings can be explained by differences in the available intervention options.

In offline settings, anonymity of bystanders is rarely discussed, but rather is conflated with discussions of responsibility (e.g., Yule and Grych, 2020). However, anonymity is potentially an important feature in online bystander decision making. In one of the few studies to use naturalistic data from 400 existing chat groups ($N = 4,833$), Markey (2000) found that simply using a bystander's name in requests for help led to a significant increase in bystander intervention compared

with a blanket request. Interestingly though, anonymity of the victim can also act as a barrier (Andalibi & Forte, 2018). Bystanders cannot see the victim's distress first hand and this may reduce the perceived seriousness of the incident. Visual anonymity of the victim can therefore create an "empathy gap" (Machackova et al., 2015), which may potentially limit intentions to help in school children.

A sense of control or belief in one's ability ("self-efficacy") is also a key enabler of helping in both online and offline scenarios (DeSmet, 2014; 2016). Song and Oh (2018) in their survey-based study with middle-school children found that a sense of control led to a small but significant effect on "defending" the victim. Similarly, Bell et al. (2019) found that Critical National Infrastructure employees were more reluctant to report concerns when they felt they did not have sufficient knowledge or skills.

Bystander intervention online: same kind of different?

On first sight, many of the same processes that underpin bystander intervention in offline settings operate in similar ways in CMC. However, it could be argued that bystander intervention online differs from offline in terms of: (a) *scale* (of victimization, of potential witnesses); (b) *timing* (i.e., victimization can be witnessed as it occurs, or some time afterward); (c) intervention *scope* (i.e., the ways in which a bystander can intervene varies online compared to offline), and (d) in *nature* (e.g., of the victim, perpetrator, act).

1) *Scale*: In a major 3-year survey of 50,000 people, Thomas et al. (2021) found that almost half (48%) reported being a target of online abuse, including being the recipient of insults or unkind comments (16%), bullying (5%), stalking (7%), and hostile account takeovers (6%). At present, we do not know if the sheer prevalence of online abuse influences users' decision making around intervention. Certainly, in some contexts, negative online behavior is seen as normative (Hilvert-Bruce & Neill, 2020), suggesting that as negative behavior in computer-mediated environments becomes more common, so the norms around what requires intervention may well change.

2) *Timing*: The nature of CMC (and, in particular, social media) means that, for many people, abuse may be witnessed sometime after it has occurred (including in screenshot format following deletion). We would expect that this might influence the likelihood of intervention because it introduces uncertainty—in terms of whether the act has been "dealt" with by others or the platform itself and indeed if an intervention would be effective. Since efficacy is a strong predictor of actual intervention (and intention to intervene) in offline settings, we would expect post hoc witnessing of abusive behavior to decrease the likelihood of intervention.

3) *Scope*: The scope for intervention in online spaces is wide, including the possibility of attempting to reduce abuse *before* it occurs (e.g., through platform rules and regulation); at the point of posting (e.g., by real-time scanning and removal of problematic content and communication); and *after* the event (e.g., through reporting and removal mechanisms). Previous research (Marwick, 2021) has shown that not all interventions are

productive, and some may even risk back-firing and amplifying the original abuse (leading to re-victimization). Some initial research in the cyberbullying literature suggests that bystanders prefer "indirect" forms of intervention (e.g., anonymous reporting) (Dillon & Bushman, 2015), presumably because it reduces the likelihood of becoming a target oneself.

4) *Nature*: Moving bystander intervention (and victimization) to a computer-mediated space allows for not only the replication of age-old methods of victimization (e.g., insults), but also new modes of abuse (Thomas et al., 2021)—what might be termed "cyber-dependent" actions (such as hacking into someone's account or zoombombing) rather than cyber-enabled re-imaginings of offline threats (such as sexual harassment, hate speech, etc.). Online environments such as social networking sites (SNS) often lack crucial social context and cues (e.g., tone of voice, distress of victim). Since ambiguity is a key barrier to intervention, we may expect that "assessing an emergency" is particularly challenging in online environments.

Theoretically, and quite naturally, the majority of work seeks to identify the prevalence and causes of abuse online. Typically, this research tends to fall into either the psychologically oriented (e.g., that trolls show evidence of Dark Triad personality traits: Craker & March, 2016), or that computer-mediated environments "enable" abuse through the affordances they provide (e.g., anonymity: Suler, 2004; Bastiaensens et al., 2015). A promising line of research has involved "designing in" (and measuring the impact of) technological interventions (e.g., notifications) with the aim of increasing bystander intervention (Bastiaensens et al., 2014; Bhandari et al., 2021; Difranzo et al., 2018; Taylor et al., 2019). While the BIM seems to apply relatively well to understanding online intervention, it is not clear how changes in the scale, timing, scope, and nature of bystander intervention within CMC might pose challenges, or extensions, to this model.

In their review of bullying (offline and online), Lambe et al. (2019) found strong support for the applicability of BIM to online settings, with 84% of the 24 reviewed studies (dated 2012–2017) finding a strong fit/correlation with the model. Of these studies, the vast majority examined school-aged children (only 5 studies looked at adults over 18) and adopted self-report methodology (18 of the 25 studies). To truly understand BIM's applicability to CMC, there is a need to interrogate each stage of the BIM in turn. Lytle et al. (2021) made a compelling start on this by proposing key differences around Step 1 (noticing) and Step 2 (assessing/ interpreting) in relation to the features that are unique to digital bystanding. However, there is still a need to elaborate on steps 3–5 of the original BIM in terms of its applicability to CMC. Furthermore, existing studies testing the applicability of the BIM have used offline intervention options (e.g., "telling a teacher") and simply applied these to online spaces (Patterson et al., 2017). Rather, we expect that there are different, perhaps more subtle ways of "intervening" within CMC (e.g., likes/shares/reposts, etc.) that have not been traditionally available offline due to the various differences (around scope, nature, and scale highlighted earlier).
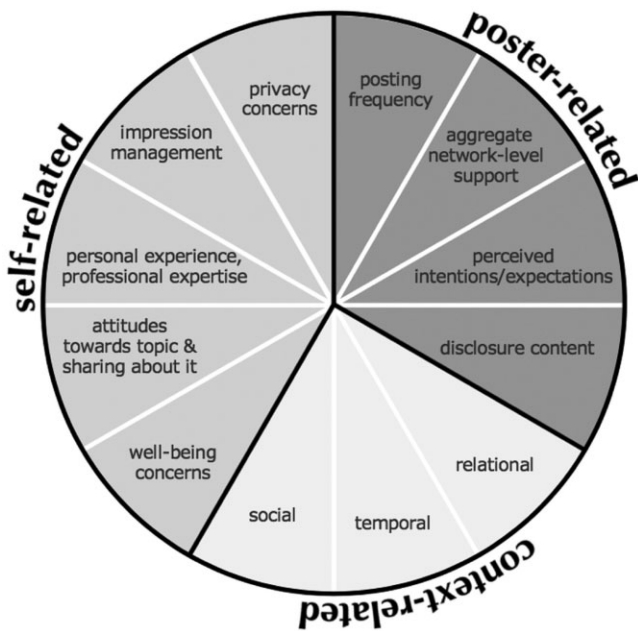
Figure 2. Response decision model (RDM) (as cited in Andalibi & Forte, 2018).

Theoretical perspective—a decision-making framework

One theoretical model that speaks to this challenge is the response decision-making framework (RDM). The RDM (see Figure 2) was developed by Andalibi and Forte (2018) to describe how users of social media respond to sensitive or socially stigmatized disclosures more broadly (e.g., mental health, pregnancy loss).

The RDM describes how various sociotechnical features (e.g., privacy settings) influence decision making. While originating from outside the traditional bystanding literature, there are many parallels to bystander intervention—specifically, RDM acknowledges how various socio-technical features can influence someone's response (e.g., an amplified sense of visibility online, sense of anonymity, the one-to-many nature of certain online postings). In particular, the RDM explains the unique pressures of responding on social media and the associated fear of social "backfiring." The framework consists of three main elements: (a) "the self" (e.g., their attitudes, impression management concerns), (b) the "poster" (e.g., type of disclosure, perceptions of their intentions), and (c) the context of the disclosure (e.g., relational, temporal, and social factors).

We assess that applying this framework of social media users' responses to socially stigmatizing material is particularly applicable to digital bystanding since it focuses specifically on the unique features of social media. Bystanders also weigh up whether (and critically, how) to respond to something concerning they see online so they may be influenced by similar factors. The present study will explore the applicability of this overarching model to digital bystanding, which has yet to be examined in this context. Taking this one step further, we propose an integration of the original BIM and the RDM into one overarching model, entitled "The Bystander Response Decision Model" (BRDM) (see Figure 3).

Through our research, we aim to pinpoint specific sociotechnical features that are relevant for each decision-making
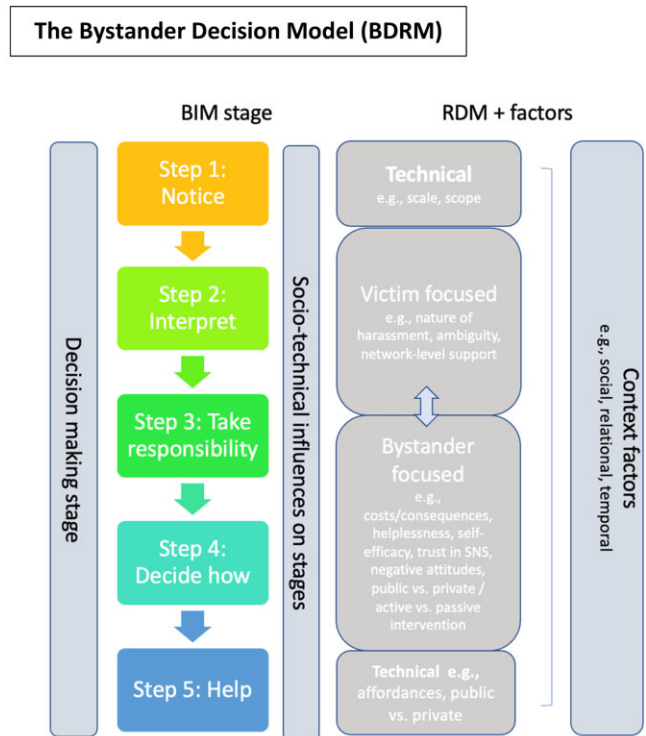
Figure 3. The BDRM—an integration of the BIM, RDM, with additional socio-technical insights from this study.

Note. BIM = bystander intervention model; RDM = response decision model.

stage (this will be discussed later). In particular, we seek to develop our understandings of what different interventions look like within CMC (e.g., stage 5 of the BIM—"deciding how to help") and possible design interventions targeted at each stage. We propose that there is a key gap in existing literature requiring further consideration. In other words, what is the full spectrum of options available within CMC, and what are the specific factors online which in turn influence decisions to select these different interventions?

Target population

In the current study, we will focus on young adult users (18–25) of social media since they are at highest risk of online harassment (UK Council for Internet Safety, 2019). As prolific users of social media platforms (Pew Research Center, 2021) we would expect young adults to witness more online harassment due to their greater exposure. Existing research has largely studied cyber-bullying in school aged populations (Barlińska et al., 2018; Brody & Vangelisti, 2015; Song & Oh, 2018) and sexual violence within University campus settings. We seek to understand the experiences of young adults in relation to online harassment on a day-to-day basis as users of SNS. Furthermore, Allison and Bussey (2016) highlight a lack of qualitative research in this field relative to scenario-based and survey designs. By taking a qualitative approach, we will develop a more nuanced understanding of the various intervention strategies used by bystanders in CMC, allowing us to unpack their decision-making process.

Research questions

The present research asks what it means to intervene in a digital world from the perspective of young adults who witness

harassment taking place. The research has three, interrelated research questions:

1) What are the barriers and enablers of helping or reporting in online contexts specifically?
2) What intervention options are available to digital bystanders, and under what circumstances are they used?
3) What are the perceived risks and benefits associated with these intervention options?

A secondary aim of the research was to identify ways of increasing online intervention to inform the way in which design and messaging can foster intervention.

Method
Design
This study adopted a qualitative methodology. As experience of online harassment is a potentially sensitive topic, focus groups were considered particularly suitable to facilitate disclosure through peer-to-peer interaction (see Guest et al., 2017). Additionally, online harassment is a social phenomenon by nature and therefore lends itself to a group-based methodology.

Participants
In total, 67 participants took part in eight focus groups with 6–10 participants in each ($N = 67$; 59.7% female, 40.3% male). A University-approved market research company recruited participants through a variety of methods (street, phone, and via email) to ensure a diverse sample and a mixture of occupations across the community. In response to a call for more research examining online harassment in young adult populations (Department of Digital, Culture, Media & Sport and Home Office, 2019), all participants were screened on two key eligibility criteria: (a) aged 18–25 years and (b) users of social media. Seven focus groups were mixed male and female participants, the eighth was all female.

Ethics
Full ethical approval was granted for the study to take place through the University of Bath's Psychology Research Ethics Committee (PREC). Several mitigations were put into place to ensure the anonymity and wellbeing of participants, and the limits of confidentiality were made clear (see the *Procedure* section). Participants provided verbal and written consent for their participation in the research and were fully aware of the aims of the research study from the outset.

Procedure
Focus groups took place in-person, before COVID-19, at hotel and conference venues in two urban locations in the UK. The focus groups were professionally facilitated by a university approved supplier with 40 years' experience. Focus groups were capped at one hour in length. Facilitators started by introducing the topic and clarifying key concepts such as a "bystander" (e.g., a witness) and "interventions" (e.g., reporting, blocking, challenging). Online harassment was described in plain terms as "harmful or abusive behavior taking place online" and deliberately kept open-ended to capture participants own views on what constitutes harassment for them.

Facilitators then checked the understanding of participants and, if necessary, provided some hypothetical scenarios (e.g., one involving a friend, one a celebrity/influencer). Given the sensitive topic of discussion, clear expectations were established at the outset about the limits of confidentiality should participants discuss criminal offences (e.g., death threats, etc.). Facilitators impressed on participants the need to report such events through formal channels. After each focus group, participants were signposted to relevant resources, including information on how to report harassment in the future. Participants were compensated with shopping vouchers (£25) for their time. The research team followed secure data handling arrangements to protect the anonymity of participants including destroying original audio recordings following transcription.

Data analysis
Focus group data were analyzed thematically following the steps outlined by Braun and Clarke (2013). Thematic analysis (TA) was considered suitable due to the complex nature of the topic and the flexible and inclusive nature of the method. TA allowed us to "…theorize motivations, experience, and meaning in a straightforward way, because a simple, largely unidirectional relationship is assumed between meaning and experience and language" (Braun & Clarke, 2006, p. 85). TA considers researcher subjectivity as a potential strength of the analysis (Braun & Clarke, 2013) and allowed us to combine diverse research backgrounds through a process of reflective practice.

Two members of the research team were responsible for the initial analysis. These researchers are from diverse research backgrounds, with differing levels of familiarity with the surrounding literature. Their different perspectives complemented one another and allowed for a more nuanced interpretation of the data. The team took steps to increase trustworthiness of their analysis, in line with Nowell et al.'s (2017) recommendations, such as having two researchers analyze the entire dataset and clearly documenting decisions and adjustments using an audit trial. A proportion of the coding was then checked by the remaining authors, plus the resulting themes sense-checked in an iterative process.

The two researchers began by individually familiarizing themselves with the data and noting initial ideas. This involved reading all focus group transcripts and noting coding ideas to inform the subsequent phases of the analysis. The transcripts were then imported into NVivo12. One researcher coded the entire dataset inductively, collated the codes, and proposed seven initial themes. The other researcher then cross-referenced their observations against the proposed coding framework in NVivo12.

The next stage of the analysis was an iterative and reflective process, where the raw data and codes were reviewed until both researchers agreed upon the finalized themes and their descriptions. At several points in this process the research team came together to triangulate ideas, discuss alternative readings of the data, refine themes, and construct mind maps.

Results
Five themes were generated from the data: Noticing and Interpreting Harassment, Perceived Responsibility for Helping, Perceived Ability to Make a Difference, Consequences of Intervening, and Deciding How to Help/

Type of Intervention. Mind maps for individual themes are included in Figures 4–7.

## Theme 1: Noticing and interpreting harassment

According to participants, the nature and severity of harassment was a fundamental factor in (a) noticing that abuse was taking place and (b) deciding to take further action. Under this overarching theme, we generated two subthemes: (a) Nature of harassment and (b) Level of ambiguity (see Figure 4).

Nature of harassment

This subtheme related specifically to: (a) immediacy and severity and (b) personal resonance. The nature (e.g., "type" of incident) and the severity of harassment influenced participants' decision to intervene. Furthermore, severe or imminent threats resulted in more immediate bystander intervention. In line with existing literature (Fischer et al., 2011; Rudnicki et al., 2022), the majority of participants felt they would immediately report incidents relating to physical safety (e.g., death threats, terrorist threats, and suicide risk). In these instances, the need for intervention was clear cut: "I think you
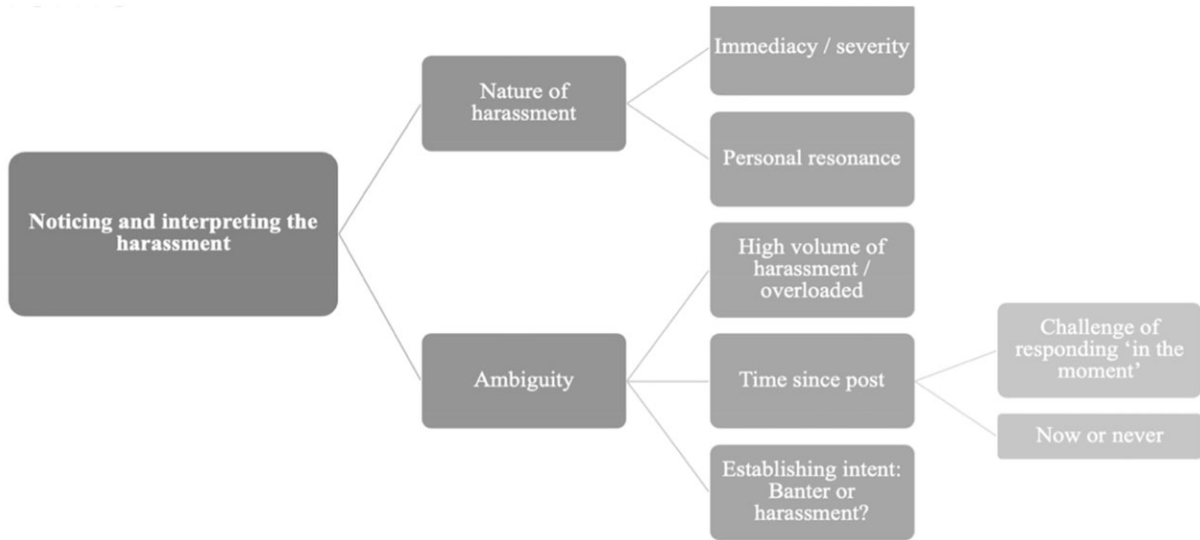
Figure 4. Thematic map for Theme 1: "Noticing and interpreting the harassment".
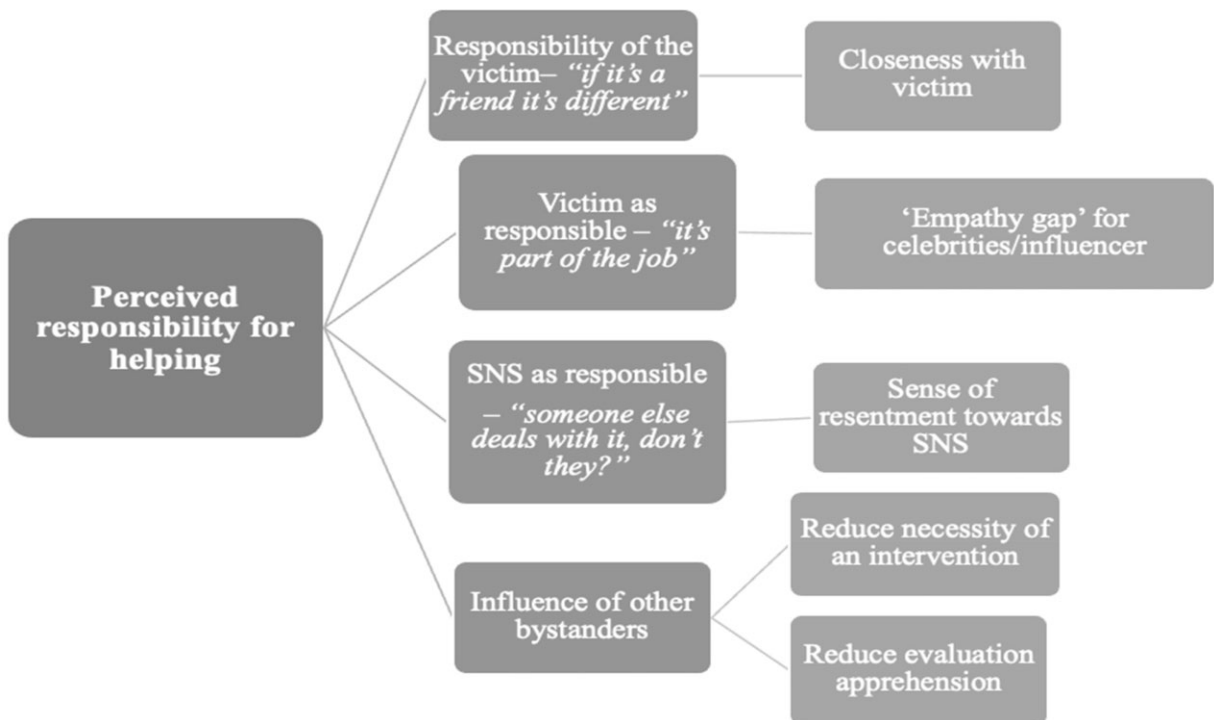
Figure 5. Thematic map for Theme 2: "Perceived responsibility for helping".
Note. SNS = social networking site.

Figure 6. Thematic map for Theme 3: "Perceived ability to make a difference".
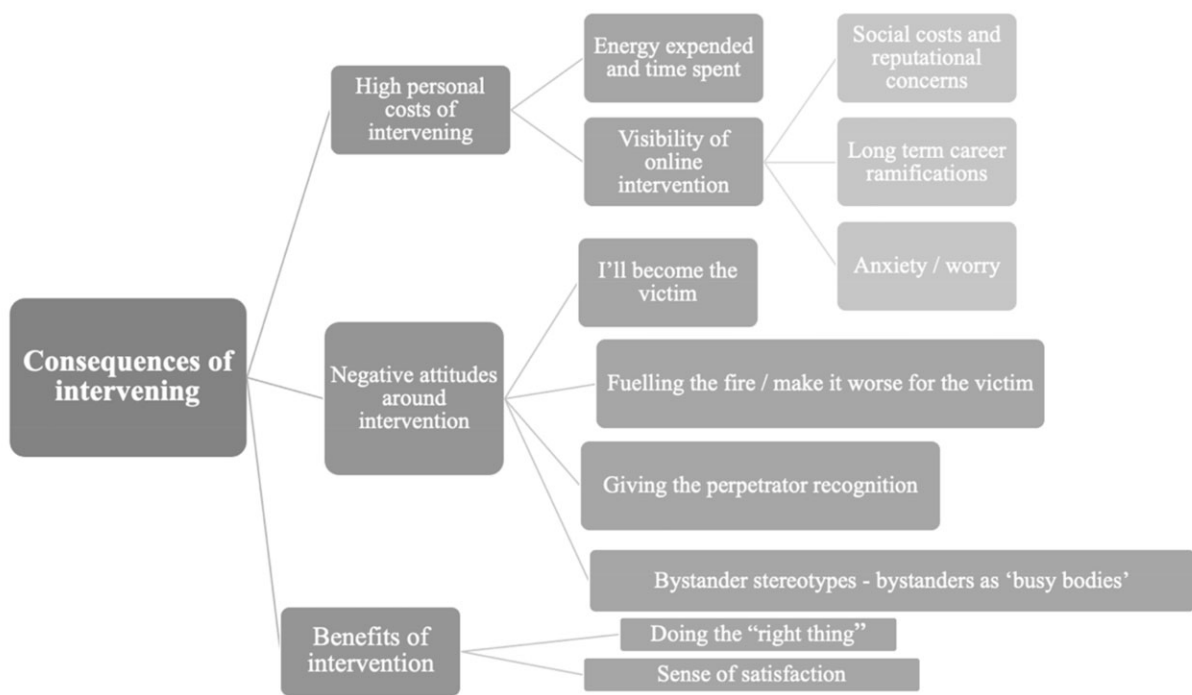

Figure 7. Thematic map for Theme 4: "Consequences of intervening".

just need to report that immediately because, like you say, it's criminal isn't it and that is immediately damaging to people" (FG 3). Participants also said they would be more likely to intervene if the abuse related to racial discrimination or sexual violence: "...you would have to step in" (FG 2). Abuse that occurs over time, such as cyber bullying, was also deemed more serious and worthy of intervention:

> ...if the abuser just posted one comment and you delete it, and they never do anything again, we can let it go. But if they constantly keep on doing it, by and by and by, then

obviously we need to report it; it becomes a serious action. (FG 6)

A few participants also commented that they would be more likely to respond to harassment that had personal resonance to them:

> It depends on the nature of things, so I guess it's what is personal to you. So, if I see anything that is slut-shaming or racism, they are the two things that for me are triggers, so I would be like, no, shut it down. (FG 3)

Ambiguity of harassment

Ambiguity related to (a) the volume of harassment, (b) time since post, and (c) difficulties establishing intent. According to Fischer et al.'s (2011) meta-analysis, individuals are much less likely to intervene if the incident is not considered urgent. In our study, the social media environment was described as highly ambiguous, making it difficult to disentangle what is an "emergency" in a traditional sense:

> It's also like when you're online, you can take stuff out of the wrong context. Like they might not mean something in some way, but because you're not face to face, because you are online, you might take it as the absolute opposite of what they meant, and then obviously that can escalate from there. (FG 4)

Participants could not physically see the reactions of the victim, which added to the ambiguity (further emphasizing the importance of body language signals; Robinson et al., 2022). Participants reported being overloaded by the sheer volume of abusive content online: "It's every day, all the time" (FG 3). Furthermore, participants found it difficult to know when an intervention was necessary: "There is a fine line of banter and then being serious. You might see it as a joke, but someone might take it too seriously. It depends on what it is" (FG 8). Ambiguity was particularly pronounced when the victim and/or abuser were not known to the bystander, making it difficult to understand the intention behind a post. "To me it didn't seem that bad...Then afterwards it was blindingly obvious I should have stepped in; I should have stopped it" (FG 5). This participant articulates the challenge of identifying online harassment in the moment.

Another important and related factor was the transitory nature of online harassment. Posts can quickly become outdated, lost, or replaced, leading to a "now or never" approach to intervention. The longer the time lag, the less likely that there is an intervention: "I am more likely to respond quite quickly or not at all" (FG 3), and the perceptions that "... I guess that issue has been dealt with" (FG 8). This time lag led some participants to assume that an intervention was not required: "If it had been three hours ago, this person's probably gone, forgotten about this, and then there's no point in replying to them" (FG 4). There was also the practical point that, "It's hard to find the same post twice on social media" (FG 3).

Theme 2: Perceived responsibility for helping

For participants to help, they needed to feel personally responsible. Under this overarching theme were four key subthemes: (a) Responsibility of the bystander, (b) Victim as responsible, (c) SNS as responsible, and (d) Influence of other bystanders (see Figure 5).

Relationship with the victim—"if it's a friend, it's different"

In line with predictions of the RDM (Andalibi & Forte, 2018), in every focus group, a commonly held view was that it was not their place to intervene when the victim was a stranger: "I guess that is the thing, isn't it. If there is enough distance almost, in a way, it is not worth your time" (FG 3). Consistent with both in-person and CMC literature (Allison & Bussey, 2016; Brody & Vangelisti, 2015; Lambe et al., 2019), when the victim was a close friend or family member,

bystanders were far more inclined to intervene in some way: "If it is someone close to me then no matter what it is, I will get involved" (FG 2). There was high consensus that "if it's a friend, it's different" (FG 3).

Further probing suggested that when the victim was a close friend, bystanders felt more able to (a) assess the likely impact of the harassment, (b) determine whether an intervention would help the victim, and (c) choose a particular intervention. For example, knowing that a friend or family member suffered from self-esteem issues helped them assess that abuse relating to physical appearance would negatively impact them, and this in turn increased their likelihood of intervening.

> I think that I would immediately react for immediate friends and family, if they attack a vulnerable point I know of them. Because every person has their own vulnerabilities, and as close friends, you know that. So, if someone says something that I know will really affect them, but if it reaches something that I know they're struggling with, there'll be definitely an immediate answer. (FG, 6)

However, this was not always the case. A small minority of participants felt that they had a moral obligation to intervene regardless of their relationship: "I feel like there's almost a duty if you're a user ...as users of the platform you should step in" (FG 7). This supports the findings of Robinson et al. (2022) that some individuals intervene for fundamental moral reasons, regardless of the situation or the relationship with the victim.

There was an interesting paradox when participants talked about the potential impact of online abuse on victims. Many participants highlighted the link between online abuse and mental health (citing high profile examples in the media such as the tragic death of Caroline Flack) and agreed that there was a need to support all victims. However, regarding celebrities or "influencers," participants consistently talked about online harassment as somewhat expected and "part of their role for being famous" (FG 6) (discussed further in theme 2— "Perceived responsibility for Helping"). It was clear that participants were aware of the potentially severe consequences of online abuse but felt that their role (as a bystander of this abuse) was limited in terms of stopping or minimizing this abuse.

SNS as responsible—"Someone else deals with it, don't they?"

Participants tended to agree that social media companies, as the original creators of these environments, should be responsible for tackling abusive behavior that occurs within them: "...the social media platform that is overseeing and allowing people to have the profiles should be accountable for letting abusive individuals onto their site" (FG 3). This finding is strikingly similar to Butler et al.'s (2022) recent work. Some recalled positive experiences of material being quickly removed from the public domain, especially when the abuse was severe. However, many participants felt that SNSs were not equipped with the right tools or resources to tackle harassment: "...I honestly feel like it's a problem that our current technological advancements are not quite capable to fix. Because like you were saying, you can't just use the robots because they're not good enough" (FG 4). Going one step further, other participants felt that SNS were not invested in

tackling the issue in the first place: "I think the social media companies need to be more responsible. They sit back with their big pay cheques and don't get involved enough. There need to be more regulations in place."

A combination of these factors led to participants feeling that the responsibility to intervene is passed onto them, as the users, instead:

> I think the reality of the situation is that we do, as users, have the responsibility, now, to be the highest responsibility…I think it should be the company's responsibility. But it isn't because they can't do that, so I think it is down to us. (FG 3)

Participants were clearly frustrated about this shift in responsibility. One participant articulated feelings of powerless: "we have no choice…they have no reason to listen to us because we use their Twitter anyway" (FG 3). It was clear that participants believed that SNSs should be doing more but lacked trust in their ability to do so.

Victim as responsible—"It's part of the job"

Some participants discussed the inevitability of online abuse, particularly in relation to celebrities and influencers where harassment was framed as an expected "part of the job" and their personal responsibility. "People should be responsible for themselves. You can't put responsibility onto other people" (FG 1). This is consistent with Weber et al.'s (2013) finding that "victim blaming" is particularly pronounced when the victim is deemed to be particularly open and extraverted online, which is often the case with "influencer" accounts. Participants mentioned examples where they felt victims had misinterpreted "banter" as targeted harassment. Taking this one step further, some participants felt that efforts to tackle this issue should be targeted at increasing the resilience of victims, rather than tackling the harassment per se.

> Maybe the best way is to train people to not be offended as much by those things [laughter]. So, I know it's hard to say, I know it's really hard, but if you train people to not be offended by some random person's comment that doesn't know you and lives somewhere… (FG 4)

Influence of other bystanders

Participants had mixed views on the influence of other bystanders, reflecting the somewhat contradictory findings we see in existing research (Allison & Bussey, 2016). However, in this study participants felt that the presence of other bystanders impacted: (a) perceived necessity of interventions and (b) perceived "evaluation apprehension." Several participants explained that it was easier to "follow the crowd" and the difficulty of being the first to intervene due to personal repercussions if the intervention backfired:

> It depends on the amount of people because it can be really hard to put that first comment out there. If it is loads of people attacking someone because they done something bad, it is easy to send a hateful tweet. To stand against that if you are by yourself is really difficult. To be the first one, it is like the Me Too movement again. (FG 2)

In contrast, others felt that the more bystanders reduced the necessity of their intervention: "If you felt that those 20 people had an impact, you could think there is no point. It is done, I don't need to get involved" (FG 5). This finding is consistent with Kazerooni et al. (2018) who found that bystanders are less likely to intervene in retweeted content vs. original content. Similarly, our participants felt an increased responsibility and moral duty to intervene when others had not, "It's like you were saying with the helping someone up in the street. If you saw it in the street and there was no-one else around, and it's only you or a couple of other people, you have to do something" (FG 2). Participants did agree, however, on the power of numbers in relation to the outcome of intervening. In fact, some participants intervened by mobilizing others to report (discussed in Theme 5: Deciding How to Help). As a rule, it was felt that the more people who reported (e.g., to SNSs), the more seriously they would be taken: "I would say the more and more people, then the report becomes more and more visible, or they start taking action probably" (FG 6).

Theme 3: Perceived ability to make a difference

A key consideration for participants was their perceived ability to make a difference. This perception of control, referred to as self-efficacy (Bandura, 1977) in the broader literature, is regularly highlighted as a key enabler of helping. In our focus group, this related to three key subthemes: (a) Knowledge, skills, and confidence, (b) lack of trust and clarity in the reporting process, and (c) sense of helplessness (see Figure 6).

Knowledge, skills, and confidence

Having knowledge of the issue or topic at hand (e.g., racial discrimination, sexual harassment) led to increased confidence in their abilities and this in turn increased their likelihood of intervening: "…if I was educated in it enough to be able to hold my own" (FG 7). The RDM framework (Andalibi & Forte, 2018) talks about the role of prior experience on the likelihood of responding to a post. The role of self-efficacy in bystander intervention is clearly acknowledged by previous research in both offline and CMC literature (Bell et al., 2019; Jouriles et al, 2018; Song & Oh, 2018). Enhancing self-efficacy is a key aim of many bystander intervention programs (Kleinsasser et al., 2015) and has been found to influence *whether* bystanders intervene and *how* they choose to intervene (DeSmet, 2014). A widely held belief was that bystanders needed to have specific skills in how to handle public online confrontations given the potential to backfire on a public scale. "I would end up making myself look silly then I would just leave myself more open to get abused then because I don't have the right thing to say" (FG 3). This again highlights the negative repercussions of helping and the preference for indirect interventions.

> If people were arguing about an issue or being abusive about something and I felt like I could hold my own in an argument about it, I might weigh in. [laughter] But if I was like I'm just going to weigh in and they're going to completely wipe the floor, [laughter] I'm just reporting that. (FG 7)

Lack of trust and clarity in the reporting process

Lack of trust in reporting related to two main factors: (a) lack of clarity around where, and how, to report online harassment in the first place and (b) reporting as an "empty" process. Most participants were not aware of the regulatory bodies or possible reporting mechanisms where they could report abusive behavior. "I think there is something that exists, I think there's two or three that exist, but I can't name them [laughter]" (FG 3). Participants felt that clearer instructions around reporting, and regular reminders, would encourage them to report in the future: "Instead of all the adverts you get, they could actually just promote how you report issues" (FG 3).

When it came to reporting abuse to SNS, participants agreed that this was often an "easy" or low effort option but were not convinced that reporting would help: "I think there is a lack of faith in the confidence of it being dealt with seriously" (FG 1). For many participants, action from the SNS did not take place quickly enough. This time lag, as mentioned previously, meant that the abuse often continued to have impact on the victim. "It is easy to report, but sometimes I feel it takes quite a while for it to be taken down. It is not always instantly …it's been over a year, and they still don't do anything?" (F6).

For other participants, reporting was described as an "empty process" (FG 4). SNSs typically responded with a generic, computerized response to acknowledge the reporting rather than a personalized message: "…reporting is the most empty thing ever. I just gave up on it because it doesn't feel like it makes a difference in any way" (FG 4). Furthermore, participants rarely knew the outcome of their reporting, and whether it helped the victim: "Now if you were to report stuff that you had no feedback on, it will stop you from reporting something" (FG 4). Several participants noted that personalized feedback would increase their confidence and, in turn, increase the likelihood of an intervention in future.

Sense of helplessness—"They are not going to listen"

A general sense of helplessness related to three main elements: (a) Inevitability of online abuse, (b) "What will I do that makes a difference?", and (3) Detachment and distance from the victim. Our focus groups suggested that when online, participants are inundated with potential opportunities to intervene. Online harassment was particularly prolific against celebrities and influencers: "You see hate comments towards people that are well known like celebrities all the time and that is in such a volume that you can't really do anything" (FG 8). As such, online harassment was seen as a common and almost inevitable feature of their daily experience.

There was a perception that they are unlikely to make a difference, regardless of how skillfully they approached an intervention:

> I picture it as you are in a crowd, people are shouting hate, you would get a microphone and say to them, "Everyone be quiet," and they will listen to you. On social media if there's loads of people giving that person hate and one person says, "Stop doing this," they are not going to listen. (FG 5)

Another participant used the powerful analogy of watching the news; it was easier to passively watch the abuse unfold rather than playing an active role as a bystander. They went on to talk about the ease of "moving on" from online harassment: "It is easier to be a bystander online than in person because you are online, you see something, flick past it, move on" (FG 5). This sense of distance was further amplified when the victim was a celebrity or online influencers: "I follow quite a lot of celebrities and people I am interested in. If I were to click on their photos there would be nine out of ten abuse on there. Do I spend my life reporting it all or do I report nothing?" (FG 4).

Theme 4: Consequences of intervening—attitudes and beliefs

Participants were typically focused on the reasons *not* to intervene and the associated risks with intervention. However, some participants did talk about a sense of satisfaction from intervention. Bystander concerns were largely organized into three main subthemes: (a) High personal costs of intervening, (b) negative attitudes around intervention, and (c) benefits of intervention (see Figure 7).

High personal costs of intervening

As with Banyard and Moynihan (2011), our participants appeared to carefully weigh the positive and negative implications of intervening. Personal costs of intervention related to (a) the energy and time expended and (b) the visibility of their intervention. The effort and time required to intervene online could be a deterrent, "it would be every single ten minutes, first it takes up some energy, but all at once, is it really useful?" (FG 4). Furthermore, participants spoke about the anxiety and fear associated with intervening online and the potential personal consequences: "If you get involved, it sucks you in and you spend your life thinking about what happened. When does it stop?" (FG 5). A widely held perception was that they could become a victim of abuse themselves if they intervened: "…I wouldn't want the hate coming towards me, because they will be giving that person hate and all the comments, and then they'd switch and it would all be on you, for the same reasons" (FG 1).

For many, the visibility of the online environment further amplified their concerns in that many potential observers could witness (and potentially judge) their intervention, a finding shared with Andalibi and Forte (2018). We know from previous literature that fear of public humiliation can be amplified online, leading bystanders to prefer offline and discreet intervention (DeSmet, 2014). Our participants elaborated on this further. "You are giving a live feed to thousands of people that are going to have 10 million different views. You are putting yourself in the firing line" (FG 2). Those who intervened (specifically, those who used public interventions) were largely viewed as "busy bodies"—people who got involved in other peoples' business: "Sometimes you're like 'Go on girl' [laughter], and other times you're like 'Oh, you must have nothing to do'" (FG 7).

In the short term, an online confrontation could lead to an increased state of anxiety for them personally until the dispute was resolved. "I think I wouldn't be able to carry on my day; I'd be thinking what are they going to say next? I'd be looking every single time [laughter], what's my next argument… I don't want to live like that" (FG 7). In the long term, due to the potential permanence and visibility of their online actions, many participants felt they could jeopardize future career

prospects. In some extreme circumstances, participants even talked about risking their personal safety if they could not remain anonymous.

Negative attitudes around intervention

The concept of "evaluation apprehension" is well established in the traditional bystander literature (Darley & Latané, 1968) and outlined in the RDM (Andalibi & Forte, 2018). As expected, participants expressed a range of anxieties and negative beliefs around intervention including: (a) I'll become the victim, (b) I'll make it worse for the victim, (c) I'll give the abuser recognition, and (d) bystanders are "busy bodies." For some, reluctance to intervene was underpinned by a sense that they would "fuel the fire" (FG 7). Elaborating on this further, participants explained that by responding to a comment, they would be shining a light on the harassment and drawing attention to it: "You feel like by ignoring it... maybe it is what you say to yourself to make yourself feel better, but you feel like you are not giving them the satisfaction of rising to it" (FG 5). For these participants, intervention could be more harmful than the original act as it can further publicize the abuse. "Scrolling past" (or privately reporting) rather than actively commenting allowed the online abuse to naturally fizzle out and disappear from the newsfeeds of others: "...if you ignore the comments or do your best to ignore the comments there is only so long someone can keep goading the same person" (FG 2).

To mitigate this, some bystanders approached the victim to ask for their consent before they publicly interviewed. For others, they felt they were "doing their bit" by not intervening and reducing the overall impact of the harassment. One participant drew on her own experience of being a victim of online harassment:

> When people started to privately message me saying, "Hope you're okay," or "I heard what's happened," that made me feel worse in a way because I know who has seen it, who's done what, what bigger platform it has been affected by. You feel more embarrassed. Because

that is one of my worst things, being embarrassed ... (FG 5)

This participant clearly illustrates the complexities and potential pitfalls around bystander intervention. There is no "one size fits all." In other words, different strategies work for different people, and some interventions have the potential to exacerbate the issue.

Benefits of intervention

A minority of participants discussed positive benefits, and these related to (a) doing the "right thing" and (b) having a personal sense of satisfaction. Participants felt that intervention can make a difference to the victim by minimizing the harm that had been caused by the abuser: "And you know that you're helping the victim. Just this feeling that by this little comment, report thing, you're doing much more than you think" (FG 6). There was also a personal sense of satisfaction and closure, "I think you know you've done your bit and you've not just ignored something that you can see is wrong. Your conscience, I suppose" (FG 7). Another participant related this back to the importance of feedback from SNSs (discussed in Theme 3), "And then if really some action is taken, then even you start feeling more confident, so there's even self-development" (FG 6). This further emphasizes the importance of providing meaningful feedback to bystanders to encourage further intervention.

Theme 5: Deciding how to help—type of intervention

In general, participants made careful and deliberate decisions regarding *how* to intervene, weighing up the potential impact of the intervention on the victim (i.e., whether it could "fuel the fire") against the personal costs of intervening. Participants referenced the following strategies, varying in: (a) level of visibility (from public to private) and (b) expended effort (from passive to active) (Figure 8).
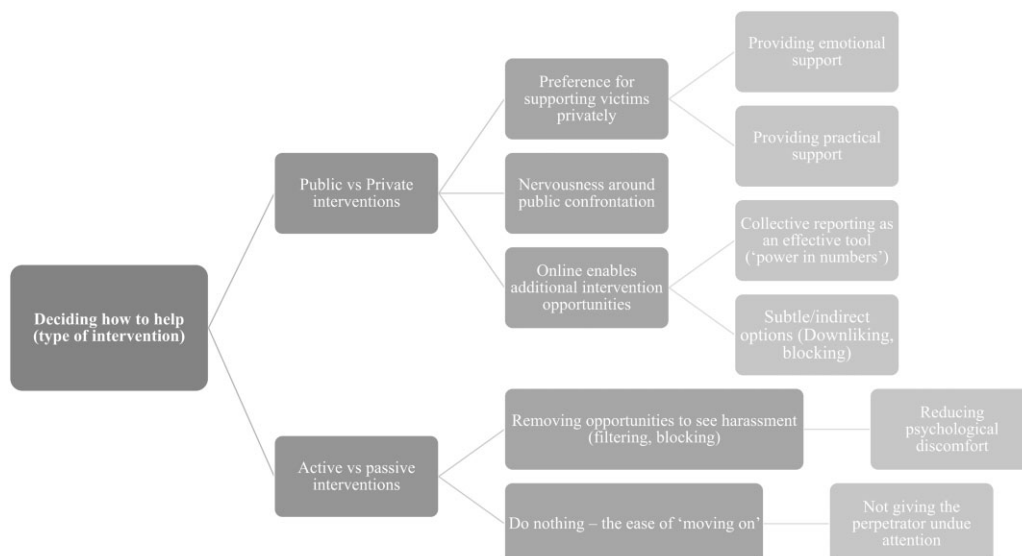

Figure 8. Thematic map for Theme 5: "Deciding how to help (type of intervention)".

Public vs. private interventions

The visibility of an intervention was a key consideration for our participants and related to: (a) preference for supporting victims privately, (b) nervousness around public confrontation, and (c) online enablers additional opportunities to intervene. As with broader CMC literature on bystanding (Difranzo et al., 2018; Dillon & Bushman, 2015), participants preferred discreet and anonymous reporting options:

> . . .because with reporting something on an anonymous basis there's no repercussions for yourself whereas if you weigh in with a comment to go straight at the abuser then they will attack you as well and it will turn into a massive escalation. (FG 3)

Very few participants felt motivated to publicly comment on the original abuse or engage in any kind of public confrontation. There was an overall sense that public interventions had higher potential consequences for them personally: "So, by responding like some of you guys have said, it's not going to change their mind, but it is then putting you in the firing line" (FG 7). As with existing research (DeSmet et al., 2014, 2016), the vast majority of participants preferred indirect interventions such as contacting the victim privately: "Depending on what it is, it is always good to reassure and have a conversation with the person affected, make sure they are okay, get the full picture of the story and be there as a friend to say, it is alright" (FG, 8). A large rationale for this was around supporting the victim and trying to reduce the overall consequences of the abuse: ". . .the victim is still going to feel bad regardless of what happens to the abuser" (FG 8). We found that these strategies aimed to "intervene" in different ways; e.g., certain strategies aimed to boost a victims' self-esteem, whereas others aimed to encourage practical action (e.g., reporting). The full range of indirect interventions is captured in Table 1.

A strong subtheme in our focus groups centered on the importance of anonymity. Participants felt much more inclined to help when their intervention would be less visible to others. For several participants, this related to a fear of public humiliation or judgement: "I don't want the embarrassment of having an argument where everyone can see it. I don't want somebody to tag me in a comment where everyone can see, and then that become visual to everybody else. I just want to keep my name quiet on it" (FG 1). In other cases, this nervousness related to the fear of retaliation from the abuser, including a fear for their personal safety.

Active vs. passive interventions

The level of involvement (and effort) from the bystander led participants to (a) remove opportunities to see harassment in the first place and (b) to "do nothing" and the ease of moving on. Reporting or blocking was a relatively quick and easy option, in comparison to more complex interventions such as providing ongoing emotional support to a victim. As discussed in Theme 3 ("Perceived Ability to Make a Difference"), participants often reported feeling helplessness, leading them to passively "watch" or scroll past harassment. Interestingly, rather than passively observing the abuse unfold, some participants removed opportunities to see abusive material in the first place (we refer to this as "avoidance strategies"). Previous research has highlighted the role of behavioral "avoidance coping strategies" in relation to cyberbullying (DeSmet et al., 2016). When asked to elaborate on this, participants described tailoring their online experience (e.g., filtering phrases on Instagram or removing known abusers from their friends list). This was summarized effectively by one participant: "It is easier to be a bystander online than in person because you are online, you see something, flick past it, move on" (FG 5).

Avoidance strategies served a particular purpose for bystanders, allowing them to protect themselves from exposure to potentially distressing material and remove moral dilemmas over whether to intervene: "And that's where, I think, in terms of hiding the comments, you would maintain your mental health … I could see it on the comments, it would really affect me. So, [laughs] then it's all about, well, it's behind doors . . ." [FG 4]. This participant alluded to the psychological discomfort that can be experienced when witnessing online abuse. Censoring, tailoring, and removing online content were considered as a potential way of alleviating this discomfort from the perspective of the bystander: ". . .so basically, if you see something, it means it's still there [laughs], and you've got the choice about whether you get involved or not" (FG 6). While not considered an "intervention" in the broader literature, these strategies impacted whether the bystander comes across abuse in the first place.

Table 1. List of "indirect" bystander intervention options with example quotes

Indirect intervention options	Example
Checking in on the victim following an incident to check their mental health	I think that would be my first port of call, just to make sure my friend's alright and see what he thinks about the situation. (FG 1)
Reporting the abuse to the social media site	And you know that you're helping the victim [by reporting]. Just this feeling that by this little comment, Report thing, you're doing much more than you think. (FG 6)
Mobilize other bystanders to report the abuse	"Yes, let's all report it and get something done about it. Six is better than one." (FG 1)
Encouraging the victim to ignore the abuse	. . . when you contact the person that is the victim, you'll say, "They're being worthless, they're being pointless, don't let it affect you, don't give it attention, don't give it your thought." (FG 4)
Get victims consent for a direct intervention, or provide practical tips	I would tell my friend to take the photo down, but also be supportive to them—but to them, not necessarily on the page (FG 8)
Boosting the victim's self-esteem/diffuse the situation	. . .maybe take a screenshot and be like "What's this person saying? As if!" Do you know what I mean? Try and make it light-hearted so, in essence, that friend isn't thinking too deeply into it. (FG 7)

Discussion

The findings in this article appear broadly consistent with the BIM (Darley & Latané, 1968) and largely mirror the original five-step model (BIM). However, our findings also highlight specific and important nuances in relation to digital bystanding that are more consistent with the RDM (Andalibi & Forte, 2018). We therefore propose merging these two models together into one overarching framework, the Bystander Decision Response Model (BRDM, see Figure 3). By doing so, we can identify specific socio-technical features relating to CMC that influence each of the five steps of the original BIM (see Table 2).

In particular, this article extends our understanding of the reasons why different micro-intervention strategies are selected and the strategic way in which they are used according to specific circumstances. In this sense, our article extends Step 4 of the original BIM ("Deciding How to Intervene") and the RDM. Only by acknowledging these nuances, which appear unique to individual platforms, can we design initiatives that target specific types of intervention. We will now highlight how many of our findings map on, and even further extend, the findings of the RDM (also see Table 2).

This study has highlighted factors that appear amplified or altered within CMC; e.g., an increased sense of ambiguity in determining what constitutes harassment. In the original five-step model, a fundamental prerequisite of helping is that an emergency is noticed in the first place and interpreted as such. Robinson et al. (2022) found that bystanders often look to "signals" (e.g., victims body language) to support their decision making—and these are not available within CMC. This begs the question of the online "signals" that digital bystanders look for when deciding whether this is indeed an emergency, clearly an area worthy of further study. In our study, ambiguity and confusion was amplified due to a saturation of information, the transitory nature of postings, and an increased sense of distance (both physically and emotionally) between bystanders and the victim. Furthermore, as per the offline bystander literature, bystanders reportedly were less likely to intervene when the situation was ambiguous, the abuse was severe (Fischer, 2006), and when there was a closer relationship with the victim (Allison & Bussey, 2016; Levine et al., 2005).

As predicted by the RDM (Andalibi & Forte, 2018), relationship closeness was a key influencer of bystander intervention. However, an important nuance should be noted around relationships and closeness in the present study. According to our participants, the closeness with the victim was less to do with sharing a group membership or common identity (as suggested in the Social Identity literature; Levine et al., 2005), and more about their ability to contextualize the harassment, reduce uncertainty, and assess the likely impact of the abuse. Knowing the victim helped bystanders gauge the type of intervention that may be most effective and allowed them to tailor their message accordingly. For some, understanding this context helped minimize the potential risk of an intervention backfiring or making the situation worse for the victim.

This study also clearly reinforces the significance of bystander feelings of efficacy, echoing findings in the sexual violence literature (Banyard & Moynihan, 2011) and cyber-bystanding literature (DeSmet et al., 2016). We also found evidence that young adults carefully balance positive and negative consequences of intervening before deciding to intervene (Banyard & Moynihan, 2011). The underlying drivers of intervention were remarkably similar to those identified by Andalibi & Forte (2018) in the RDM. In particular, a sense of having the right skills or expertise to respond (e.g., professional background or personal experience of the issue at hand) was a key enabler or barrier to intervention. Furthermore, a sense of being judged or misunderstood by others was a key impression management concerns—in both the RDM and our current study (Andalibi & Forte, 2018). In our study, participants also considered whether they were the first to respond or whether the harassment had been reshared; as was the case in Kazerooni et al. (2018).

In terms of taking responsibility for helping (Step 3 of the BIM) participants often felt it was the responsibility of platforms rather than users to address online harassment, a finding shared with Butler et al. (2022). This finding has important implications for SNS as it suggests that by advertising/promoting their tools they may actually be disempowering users from tackling harassment at the grassroots. In our study, the perceived costs of intervening appeared to be further amplified through CMC, and often vastly outweighed positive benefits of intervening. We found that young adults had a significant fear of repercussions (i.e., that there would be significant personal, safety, or social costs for them) due to the public and "permanent" nature of SNS. These findings mirror those outlined in the RDM, in particular "self-related" concerns such as impression management concerns, level of privacy, and their own personal well-being (Andalibi & Forte, 2018). They were also anxious that they would unwittingly exacerbate the abuse for the victim (i.e., "fuel the fire") if their intervention backfired, or that they could become a victim themselves, a finding that is also reflected in Marwick's (2021) qualitative work on harassment whereby other users can inadvertently perpetuate the abuse. Our findings highlight the need for SNS to equip young adults with not only the tools (e.g., quick and easy reporting options) but the skills and confidence in order to intervene. The benefits of intervention need to be made clearer to individual users, highlighting real examples of where a bystanders' behavior has stopped or reduced harm.

Another striking finding in the current study was the range of potential interventions that bystanders can take online (see Table 1 and Figure 9): varying in complexity, effort/risk, and level of visibility. As with the RDM (Andalibi & Forte, 2018), participants often used different levels of privacy (e.g., "liking" rather than commenting) to alleviate some of their concerns (e.g., impression management concerns). Not all these interventions had an "offline" equivalent (e.g., "downliking" an abuser's comments, for example, to show a lack of support), perhaps suggesting that online environments can offer new and innovative ways for bystanders to intervene. Mirroring findings in the cyber-bullying literature, we also found a strong preference for "indirect interventions" (see Table 1), such as reporting the abuse to SNS or supporting victims (Dillon & Bushman, 2015; Fischer et al., 2011). However, we found that online bystanders often remove and avoid opportunities to notice the abuse in the first place (e.g., tailoring and censoring their newsfeed). According to the original BIM, awareness of an incident is a prerequisite of bystander intervention (Darley & Latané, 1968). Such avoidance behaviors could be described as coping strategies rather than active interventions, to deal with the high volume of online harassment that young adults have become so

Table 2. A summary of key predictions of the RDM for each stage of the BIM

Decision making process (BIM)	Poster/victim related (RDM)	Self/bystander related (RDM)	Context related (RDM)	Additional CMC factors (this study)	Possible design interventions
Step 1: Notice that something is happening	/	/	/	• Ambiguity • Nature of harassment (e.g., volume, intent, time since post severity)	• More guidance/training around what constitutes harassment within CMC
Step 2: Interpret as an emergency (/harassment)	• Posting frequency • Network-level support • Perceived intentions/ Expectations • Original message content		• Relational (closeness) • Social (e.g., norms) • Temporal (necessity)	• Ambiguity • Nature of harassment (e.g., volume, intent, time since post)	• More guidance/training around what constitutes harassment within CMC
Step 3: Take responsibility for helping		• Attitudes toward topic • Personal experience, professional expertise	• Relational (closeness) • Social (e.g., norms) • Temporal (necessity)	• SNS as responsible • Influence of other bystanders • Victim as responsible (e.g., influential users)	• Reminder of the limits of SNS moderation and the responsibilities of individual users • Mandatory bystander training prior to sign up. Exposure to cases studies of harassment to increase sense of personal responsibility • Feedback on the impact on victims (e.g., visual cue that victim has been offended)
Step 4: Decide how to help		• Personal experience, professional expertise • Well-being concerns • Impression management • Privacy concerns	• Relational (closeness) • Social (e.g., norms) • Temporal (necessity)	• Lack of trust in reporting process • Sense of helplessness • Negative attitudes around intervention • High personal costs of intervening • Benefits of intervention	• Efficacy nudges e.g., reminding users of intervention options through pop-up messaging, clearer signposting to reporting options • Incentivize intervention. SNS to provide feedback to encourage future reporting
Step 5: Provide help	/	/	/	• Active vs. passive • Public vs. private See "spectrum of interventions" (Figure 9)	• Clear, easy, and anonymous reporting options to reduce privacy concerns • More "low risk" options (e.g., down-liking) across all SNS • Easy collective reporting options • Education on more indirect strategies that can be used

Note. In addition, this highlights additional socio-technical features that were identified by the current research.
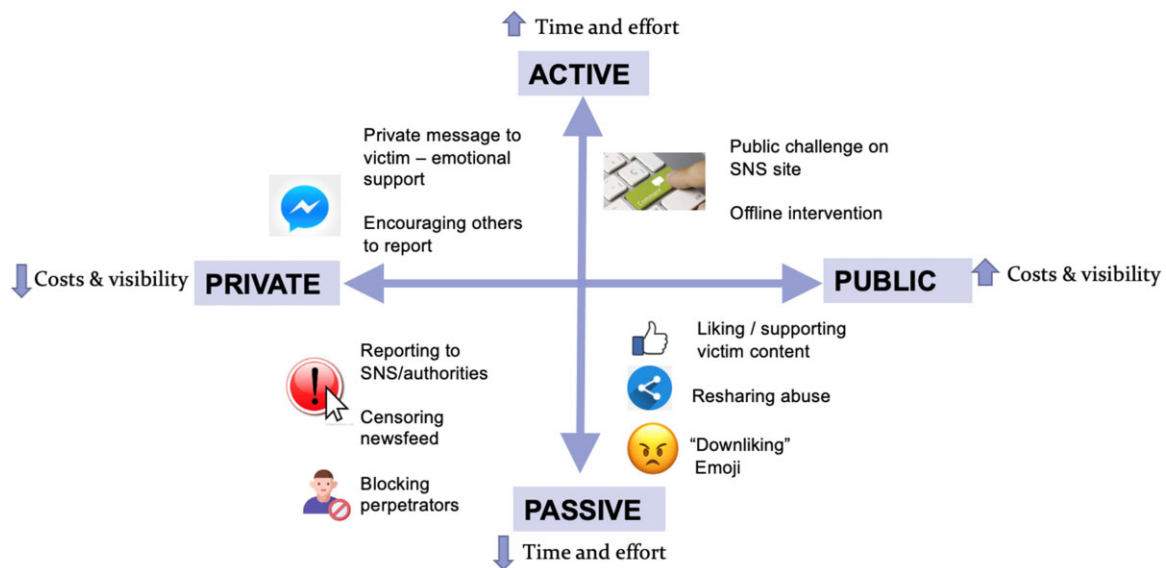
Figure 9. A conceptualization of bystander intervention types within CMC.
Note. CMC = computer-mediated communication; SNS = social networking site.

accustomed to. We therefore suggest a move away from binary conceptualizations of bystander intervention toward a continuum made up of a spectrum of behaviors (see Figure 9). According to participants, SNS and bystander initiatives could do more to remind users of the various intervention strategies along this spectrum.

Taken together, these findings emphasize the importance of bystander training and education in equipping young adults with the relevant skills, confidence, and resources to intervene. These campaigns should consider the variety of ways in which digital bystanders can intervene within CMC. In addition, users need to understand the potential value of intervening through successful case studies, and by providing them with specific feedback on how their intervention helped. This finding again reinforces the importance of bystander training and education that is specifically tailored to mediated online environments, taking into account the barriers that young adults experience.

Limitations

It is important to acknowledge some fundamental limitations to the current study. First, as with focus group research of this nature, participants may have felt the need to converge their opinions, and this may have impacted the extent to which participants could candidly discuss their views. While we made active attempts to minimize this through experienced mediation, it is possible that some participants censored their views. Another limitation is the self-report nature of the data and the lack of ground truth data (i.e., actual intervention). Specifically, there is a need to test whether the messaging, or design, of SNSs can lead to an increase in actual intervention. Another productive line of research would be to look at the relative success of various online intervention strategies investigating the notion of whether intervention is always the "right" solution.

Future research and applications

By understanding the decision-making process underlying digital bystander behavior, we are another step forward in terms of reducing the blockers and incentivizing intervention. We argue that taken together, the RDM and the BIM offer a more comprehensive model that can help us understand key drivers for digital bystanding. BIM offers a high-level sequential pathway underlying decision making, whereas RDM breaks down the socio-technical influences on these decision points. Through this research, we connect these two, traditionally separate models, into one overarching model (BDRM—Figure 3 and Table 2). We have identified the features of CMC that are relevant to different stages of the BIM and believe that this forms the basis for designing targeted and effective design interventions. Furthermore, we have identified a "spectrum" of intervention options that are available online—and propose that simply applying traditional "offline" models to describe different types of intervention is no longer sufficient. We now encourage further testing and development of these models within the context of digital bystanding.

According to our participants, there is a need to educate (and continually remind) users about the severe and potentially fatal consequences of online harassment, letting people know the effects of that, circumstances, consequences, what could potentially happen…" (FG 8). The recent tragic death of a celebrity was mentioned by several participants as a stark reminder of the possible implications of online harassment. Participants impressed the need to educate users on what online harassment can look like and to encourage SNSs to take more responsibility.

The current article emphasizes the importance of "bystander friendly" design features, such as the role of anonymous reporting options. One participant suggested the equivalent of a "stop, look, listen" message to be instilled in users before posting. Being able to control and tailor their online experience was extremely important in terms of managing online harassment. This study has highlighted variables that could be "designed in," building on the growing number of studies in the field (Difranzo et al., 2018., Dillon & Bushman, 2015). For example, the research highlights the importance of feedback in creating a transparent and trustworthy reporting process.

Conclusion

Given the increased volume and severity of online harassment and the limitations of automated tools (Rosa et al., 2019), there is an urgent need to focus efforts on human users, or bystanders. This study suggests that an interplay of socio-technical factors underpin decisions to intervene and influence the way in which bystanders intervene (see proposed model "BRDM," Figure 3). Traditional bystander models are relevant; however, this study highlights additional factors that are unique to CMC and therefore essential to incorporate into the design of SNSs. Building on early research (Difranzo et al., 2018; Dillon & Bushman, 2015), we call for CMC researchers to empirically test these qualitative findings through design and messaging, and for SNSs to implement these features to incentivize different types of intervention.

Data availability

The data underlying this article cannot be shared publicly for the privacy of individuals that participated in the study and due to the sensitive nature of this topic. Anonymized themes may be shared on reasonable request to the corresponding author.

Funding

References

Andalibi, N., & Forte, A. (2018). Responding to sensitive disclosures on social media: A decision-making framework. *ACM Transactions on Computer–Human Interaction*, 25(6), 1–29. https://doi.org/10.1145/3241044

Anderson, J., Bresnahan, M., & Musatics, C. (2014). Combating weight-based cyberbullying on Facebook with the dissenter effect. *Cyberpsychology, Behavior and Social Networking*, 17(5), 281–286. https://doi.org/10.1089/cyber.2013.0370

Allison, K. R., & Bussey, K. (2016). Cyber-bystanding in context. A review of the literature on witnesses' responses to cyberbullying. *Children & Youth Service*, 65, 183–194. https://doi.org/10.1016/j.childyouth.2016.03.026

Barlińska, J., Szuster, A., & Winiewski, M. (2018). Cyberbullying among adolescent bystanders: Role of affective versus cognitive empathy in increasing prosocial cyberbystander behavior. *Frontiers in Psychology*, 9, 1–13. https://doi.org/10.3389/fpsyg.2018.00799

Bastiaensens, S., Vandebosch, H., Poels, K., Van Cleemput, K., DeSmet, A., & de Bourdeaudhuij, I. (2015). 'Can I afford to help?' How affordances of communication modalities guide bystanders' helping intentions towards harassment on social network sites. *Behaviour & Information Technology*, 34(4), 425–435. https://doi.org/10.1080/0144929X.2014.983979

Bastiaensens, S., Vandebosch, H., Poels, K., Cleemput, K., DeSmet, A., & Bourdeaudhuij, I. (2014). Cyberbullying on social network sites. An experimental study into bystanders' behavioural intentions to help the victim or reinforce the bully. *Computers in Human Behavior*, 34, 259. https://doi.org/10.1016/j.chb.2013.10.036

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioural change. *Psychological Review*, 84(2), 191–215. https://doi.org/10.1037/0033-295X.84.2.191

Banyard, V. L., & Moynihan, M. M. (2011). Variation in bystander behavior related to sexual and intimate partner violence prevention: Correlates in a sample of college students. *Psychology of Violence*, 1(4), 287–301. https://doi.org/10.1037/a0023544

Bell, A. J. C., Rogers, M. B., & Pearce, J. M. (2019). The insider threat: Behavioural indicators and factors influencing likelihood of intervention. *International Journal of Critical Infrastructure Protection*, 24, 166–176. https://doi.org/10.1016/j.ijcip.2018.12.001

Bhandari, A., Ozanne, M., Bazarova, N. N., & DiFranzo, D. (2021). Do you care who flagged this post? Effects of moderator visibility on bystander behavior. *Journal of Computer-Mediated Communication*, 26(5), 284–300. https://doi.org/10.1093/jcmc/zmab007

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3, 77–101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2013). *Successful qualitative research: A practical guide for beginners*. SAGE Publications.

Brody, N., & Vangelisti, L. A. (2015). Bystander intervention in cyberbullying. *Communication Monographs*, 83(1), 94–119. https://doi.org/10.1080/03637751.2015.1044256

Brody, N. (2013). *Bystander intervention in cyberbullying* [Unpublished doctoral dissertation]. The University of Texas at Austin. http://hdl.handle.net/2152/21636

Butler, L. C., Graham, A., Fisher, B. S., Henson, B., & Reyns, B. W. (2022). Examining the effect of perceived responsibility on online bystander intervention, target hardening, and inaction. *Journal of Interpersonal Violence*, 37, 21–22. https://doi.org/10.1177/08862605211055088

Craker, N., & March, E. (2016). The dark side of Facebook: The Dark Tetrad, negative social potency, and trolling behaviours. *Personality and Individual Differences*, 102, 79–84. https://doi.org/10.1016/j.paid.2016.06.043

Department of Digital, Culture, Media & Sport and Home Office. (2019). *Online harms white paper*. Her Majesty's Stationary Office. www.gov.uk/government/publications

Darley, J. M., & Latané, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology*, 8(4), 377–383. https://doi.org/10.1037/h0025589

DeSmet, A., Veldeman, C., Poels, K., Bastiaensens, S., Van Cleemput, K., Vandebosch, H., & De Bourdeaudhuij, I. (2014). Determinants of self-reported bystander behavior in cyberbullying incidents among adolescents. *Cyberpsychology, Behavior and Social Networking*, 17(4), 207–215. http://dx.doi.org/10.1089/cyber.2013.0027

DeSmet, A., Bastiaensens, S., Van Cleemput, K., Poels, K., Vandebosch, H., Cardon, G., & De Bourdeaudhuij, I. (2016). Deciding whether to look after them, to like it, or leave it: A multidimensional analysis of predictors of positive and negative bystander behavior in cyberbullying among adolescents. *Computers in Human Behavior*, 57, 398–415. https://doi.org/10.1016/j.chb.2015.12.051

Difranzo, D., Taylor, S. H., Kazerooni, F., Wherry, O. D., & Bazarova, N. N. (2018). *Upstanding by design: Bystander intervention in cyberbullying*. Paper presented at the CHI Conference on Human Factors in Computers Systems, Montreal, QC, Canada. https://doi.org/10.1145/3173574.3173785

Dillon, K. P., & Bushman, B. J. (2015). Unresponsive or un-noticed?: Cyberbystander intervention in an experimental cyberbullying context. *Computers in Human Behavior*, 45, 144–150. https://doi.org/10.1016/j.chb.2014.12.009

Guest, G., Namey, E., Taylor, J., Eley, N., & McKenna, K. (2017). Comparing focus groups and individual interviews: Findings from a randomized study. *International Journal of Social Research Methodology*, 20(6), 693–708. https://doi.org/10.1080/13645579.2017.1281601

Fenton, R. A., Mott, H. L., McCartan, K., & Rumney, P. (2016) *A review of evidence for bystander intervention to prevent sexual and domestic violence in universities*. Technical Report. Public Health England. https://www.researchgate.net/publication/341165468_A_review_of_evidence_for_bystander_intervention_to_prevent_sexual_and_domestic_violence_in_universities

Fischer, P. (2006). The unresponsive bystander: Are bystanders more responsive in dangerous emergencies? *European Journal of Social Psychology*, 36(2), 267–278. https://doi.org/10.1002/ejsp.297

Fischer, P., Krueger, J. I., Greitemeyer, T., Vogrincic, C., Kastenmüller, A., Frey, D., Heene, M., Wicher, M., & Kainbacher, M. (2011). The bystander-effect: A meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. *Psychological Bulletin*, 137(4), 517–537. https://doi.org/10.1037/a0023304

Hilvert-Bruce, Z., & Neill, T. J. (2020). I'm just trolling: The role of normative beliefs in aggressive behaviour in online gaming. *Computers in Human Behavior*, 102, 303–311. https://doi.org/10.1016/j.chb.2019.09.003

Jouriles, A. N., Krauss, A., Vu, N. L., Banyard, V. L., & McDonald, R. (2018). Bystander programs addressing sexual violence on college campuses: A systematic review and meta-analysis of program outcomes and delivery methods, *Journal of American College Health*, 66(6), 457–466. https://psycnet.apa.org/doi/10.1037/vio0000284

Katz, J., & Nguyen, L J. (2016). Bystander responses to risk for rape perpetrated by a friend, acquaintance, or stranger. *Journal of Aggression, Maltreatment & Trauma*, 25(6), 652–667. https://doi.org/10.1080/10926771.2016.1185755

Kazerooni, F., Taylor, S. H., Bazarova, N. N., & Whitlock, J. (2018). Cyberbullying bystander intervention: The number of offenders and retweeting predict likelihood of helping a cyberbullying victim. *Journal of Computer-Mediated Communication*, 23(3), 146–162. https://doi.org/10.1093/jcmc/zmy005

Kiesler, S., Siegel, J., & McGuire, T. (1984). Social psychological aspects of computer-mediated communication. *American Psychologist*, 39(1), 1123–1134. https://doi.org/10.1037/0003-066X.39.10.1123

Kleinsasser, A., Jouriles, E. N., McDonald, R., & Rosenfield, D. (2015). An online bystander intervention program for the prevention of sexual violence. *Psychology of Violence*, 5(3), 227–235. https://doi.org/10.1037/a0037393

Lambe, L. J., Della Cioppa, V., Hong, I. K., & Craig, W. M. (2019). Standing up to bullying: A social ecological review of peer defending in offline and online contexts. *Aggression and Violent Behavior*, 45, 51–74. https://doi.org/10.1016/j.avb.2018.05.007

Latané, B., & Darley, J. M. (1970). *The unresponsive bystander*. Appleton-Century-Croft. New York, NY.

Levine, M., Prosser, A., Evans, D., & Reicher, S. (2005). Identity and emergency intervention: How social group membership and inclusiveness of group boundaries shape helping behavior. *Personality and Social Psychology Bulletin*, 31(4), 443–453. https://doi.org/10.1177/0146167204271651

Liu, C., Liu, Z., & Yuan, G. (2020). The longitudinal influence of cyberbullying victimization on depression and posttraumatic stress symptoms: The mediation role of rumination. *Archives of Psychiatric Nursing* 34(4), 206–210. https://doi.org/10.1016/j.apnu.2020.05.002

Lytle, R. D., Bratton, T. M., & Hudson, H. K. (2021). Bystander apathy and intervention in the era of social media. In J. Bailey, A. Flynn, & N. Henry (Eds.), *The emerald international handbook of technology-facilitated violence and abuse* (pp. 711–728). Emerald Publishing Limited. https://doi.org/10.1108/978-1-83982-848-520211052

Machackova, H., Dedkova, L., & Mezulanikova, K. (2015). Brief report: The bystander effect in cyberbullying incidents. *Journal of Adolescence*, 43, 96–99. https://doi.org/10.1016/j.adolescence.2015.05.010

Marwick, A. E. (2021). Morally motivated networked harassment as normative reinforcement. *Social Media + Society*, 7(2). https://doi.org/10.1177/20563051211021378

Markey, P. M. (2000). Bystander intervention in computer-mediated communication. *Computers in Human Behavior*, 16(2), 183–188. https://doi:10.1016/S0747-563(99)00056-4

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness Criteria. *International Journal of Qualitative Methods*, 16(1), 1–21. https://doi.org/10.1177/1609406917733847

Obermaier, M., Fawzi, N., & Koch, T. (2016). Bystanding or standing by? How the number of bystanders affects the intention to intervene in cyberbullying. *New Media & Society*, 18(8), 1491–1507. https://doi.org/10.1177/1461444814563519.

Patterson, L. J., Allan, A., & Cross, D. (2017). Adolescent bystander behavior in the school and online environments and the implications for interventions targeting cyberbullying. *Journal of School Violence*, 16(4), 361–375. https://doi.org/10.1080/15388220.2016.1143835

Perren, S., Dooley, J., Shaw, T., & Cross, D. (2010). Bullying in school and cyberspace: Associations with depressive symptoms in Swiss and Australian adolescents. *Child and Adolescent Psychiatry and Mental Health*, 4(28), 1–10. https://doi.org/10.1186/1753-2000-4-28

Pew Research Center. (2021). *Online harassment survey*. https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/

Pew Research Center. (2017). *Online harassment survey*. https://www.pewinternet.org/2017/07/11/online-harassment-2017/

Robinson, S. R., Casiano, A., & Elias-Lambert, N. (2022). "Is it my responsibility?": A qualitative review of university students' perspectives on bystander behavior. *Trauma, Violence, & Abuse*, 23(1), 117–131. https://doi.org/10.1177/1524838020933829

Rosa, H., Pereira, N., Ribeiro, R., Ferreira, P. C., Carvalho, J. P., Oliveira, S., & Trancoso, I. (2019). Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior*, 93, 333–345. https://doi.org/10.1016/j.chb.2018.12.021

Rudnicki, K., Ouvrein, G., De Backer, C., & Heidi, V. (2020). Intranasal oxytocin administration reduces bystanders' acceptance of online celebrity bashing. *International Journal of Bullying Prevention*, 2(1), 29–40. https://doi.org/10.1007/s42380-019-00054-2

Rudnicki K., Vandebosch, H., Voué, P., & Poels, K. (2022). Systematic review of determinants and consequences of bystander interventions in online hate and cyberbullying among adults. *Behaviour & Information Technology*, 42(5), 527–544. https://doi.org/10.1080/0144929X.2022.2027013

Song, J., & Oh, I. (2018). Factors influencing bystanders' behavioural reactions in cyberbullying situations. *Computers in Human Behavior*, 78, 273–282, https://doi:10.1016/j.chb.2017.10.008

Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & behavior : the impact of the Internet, multimedia and virtual reality on behavior and society*, 7(3), 321–6. https://doi.org/10.1089/1094931041291295.15257832

Taylor, S., DiFranzo, D., Choi, Y., Sannon, S., & Bazarova, N. (2019). Accountability and empathy by design. *Proceedings of the ACM on Human–Computer Interaction*, 3(CSCW), 1–26. https://doi.org/10.1145/3359220

Thomas, K., Akhawe, D., Bailey, M., Boneh, D., Bursztein, E., Consolvo, S., Dell, N., Durumeric, Z., Kelley, P., Kuma, D., McCoy, D., Meiklejohn, S., Ristenpart, T., & Stringhini, G. (2021). SoK: Hate, harassment, and the changing landscape of online abuse. In *2021 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA* (pp. 247–267). https://doi: 10.1109/SP40001.2021.00028

UK Council for Internet Safety. (2019). Adult online hate, harassment, and abuse: A rapid evidence assessment. https://www.gov.uk/government/publications/adult-online-hate-harassment-and-abuse-a-rapid-evidence-assessment

Van Cleemput, K., Vandebosch, H., & Pabian, S. (2014). Personal characteristics and contextual factors that determine "helping," "joining in," and "doing nothing" when witnessing cyberbullying. *Aggressive Behavior*, 40(5), 383–396. https://doi.org/10.1002/ab.21534

Wang, S. (2021). Standing up or standing by: Bystander intervention in cyberbullying on social media. *New Media & Society*, 23(6), 1379–1397. https://doi.org/10.1177/1461444820902541

Weber, M., Ziegele, M., & Schnauber, A. (2013). Blaming the victim: The effects of extraversion and information disclosure on guilt attributions in cyberbullying. *Cyberpsychology, Behavior, and Social Networking*, 16, 254–259. https://doi.org/10.1089/cyber.2012.0328

Yule, K., & Grych, J. (2020). College students' perceptions of barriers to bystander intervention. *Journal of Interpersonal Violence*, 35(15–16), 2971–2992. https://doi.org/10.1177/0886260517706764