Unmasking Anonymity: The Effect of Aggressor Anonymity in Online Hate Speech on Bystanders' Intention to Intervene

Belle de Folter

SNR 2123312

Master's Thesis

Communication and Information Sciences

Specialization Business communication and Digital Media

Department Communication and Cognition

School of Humanities and Digital Sciences, Tilburg University

Tilburg University, Tilburg

Supervisor: Dr. J. Verhoeven

Second reader: Dr. S.J.R. Pabian

Januari 2025

Technology statement

While authoring the thesis, I used Quillbot as a tool to help paraphrase. Quillbot, Grammarly and Word were used to check for grammar and spelling mistakes, and for some paragraphs to improve the flow and readability. I used Perplexity, SciSpace, and Consensus.ai in addition to academic source banks to detect some sources. However, I did check all sources to ensure that these AI tools did not make mistakes. I did not use ChatGPT for the detection of these studies and findings, or to generate complete paragraphs. Finally, I used Scribbr for the storage of the references.

By submitting my thesis for assessment, I hereby confirm that:

- the thesis is my own intellectual property and that ideas as well as language from other sources have been properly cited. All quotes and sourced information have been properly identifiable as such.
- I have disclosed any technology that I have used in the writing process.

Abstract

This study investigates the role of bystanders in diminishing online hate and explores how and whether aggressor anonymity influences their intention to intervene. Bystanders could play a crucial role in the establishment of clear behavior guidelines by means of social sanctioning. Using a between-subjects experimental design, 179 participants were exposed to hate speech comments from aggressors with varying levels of anonymity: identifiable, pseudonymous, and anonymous. They were asked whether they intended to intervene through counterspeech (direct intervention) or flagging or reporting (indirect intervention). Mediation analyses (Hayes Models 4 and 6) were employed to examine the mediating role of perceived social presence, credibility, and threat in shaping direct or indirect intervention intentions. Results revealed that none of the hypothesised relationships were significant, except for the indirect relationship between aggressor anonymity, perceived threat, and indirect intervention intention. Anonymous aggressors were perceived as more threatening, which indirectly increased intentions for indirect interventions such as reporting or flagging content. These low indirect interventions intention suggest that one can not rely on bystanders to actively engage in in counterspeech. Relying on bystanders as norm reinforcers is therefore an ineffective and overly idealistic approach. Instead, focusing on bystanders as frontline detectors of online hate speech represents a more realistic solution. Focusing on indirect actions and enhancing digital literacy can bridge the gap between human and automated efforts to effectively combat online hate speech.

Unmasking Anonymity: The Effect of Aggressor Anonymity in Online Hate Speech on Bystanders' Intention to Intervene

In the last decade, the conversational tone on social networking sites (SNS) has become increasingly hostile, making exposure to hate speech almost unavoidable (Schmid et al., 2022). Hate speech includes any form of communication that disparages, encourages, incites, or expresses hatred or incivility toward an individual or group based on specific characteristics, such as race, color, ethnicity, gender, sexual orientation, national origin, or religion (Salminen et al., 2020; Schäfer et al., 2022; Siahaan, 2023). Scholars suggest that the prevalence of hate speech has been significantly influenced by online anonymity, as it enables users to express more offensive opinions without fearing social or legal repercussions (Qureshi & Sabih, 2021; Wang, 2020; Woods & Ruscher, 2021). Hateful content spreads faster, and more widely and reaches a wider audience, increasing the prevalence of online hate speech (Mathew et al., 2019). In fact, Castaño-Pulgarín et al. (2021) suggest that 80% of European Union citizens have been exposed to online hate speech, and 40% have been targeted (directly or indirectly) by online hate speech aggressors on SNS.

Online hate speech harms not only individuals directly targeted but also impacts marginalized groups and society as a whole. According to Quirk & Campbell (2014), individuals in hate speech scenarios may assume one of three roles: victim, aggressor, or bystander. The effects of hate speech are often studied in the context of direct victims or vicarious victims, suggesting that they could both experience short-term effects ranging from shock, loneliness, or rage to long-term behavioral effects like social exclusion or increased mistrust in strangers (Agarwal et al., 2021; Leonhard et al., 2018). Moreover, user comment sections serve as exemplars for society and influences readers' opinions and attitudes towards the world (Schäfer, 2022). As a result, bystanders, irrespective of their behavior, represent the social consensus, with their actions or silence upholding or challenging social norms (Craig et

al., 2000). If this hateful tone maintains unchallenged, this may eventually lead to uninvolved bystanders developing warped perceptions of hate speech, further reducing their intention to criticize such remarks and change their worldviews even more (Leonhard et al., 2018).

Bureaucratic control methods, defined as the employment of regulations to establish control over its community (Rennstam, 2017), have been ineffective in the removal of hate speech on SNS (Wilson & Land, 2021). Studies on natural language processing technology explored the possibility of automatic textual hate speech detection in recent years (Jahan & Oussalah, 2023; Salminen et al., 2020; Zhang & Luo, 2019). However, scholars argue that restricting free speech creates a disturbing precedent for selective expression (Mathew et al., 2018; Rashidi et al., 2020; Wang et al., 2020). Obermaier (2022) suggests that counterspeech from bystanders most effectively reduces online hate speech because of its visibility in comment sections. This direct disapproval sanctions the hate speech while also upholding the principles of free speech. In this process of social sanctioning, members learn from other members of a group or community what behavior is approved and what behavior will be sanctioned (Álvarez-Benjumea & Winter, 2018). Counterarguing is a form of normative social control and is argued to be more effective than bureaucratic control (Rennstam, 2017). Indirect measures, such as blocking or automatically deleting hate speech, are ineffective on a large scale. Their lack of direct disapproval from the group may lead violators to interpret silence as consent and continue spreading hate speech (Matthew et al., 2018; Rashidi et al., 2020).

SNS enable individuals to freely engage in arguments and criticism without the constraints of physical presence, traceability, or identifiability (Culpeper, 1996). This option to remain anonymous is defined as the degree to which a communicator perceives the message source as unknown and unspecified (Cho et al., 2012). However, according to Marx (1999), we live in a culture where identification is the norm, and remaining anonymous can

conflict with cultural expectations. Although anonymity enables open communication, it is still expected that basic social norms, like politeness and respect, will be followed. Hate speech directly violates these norms and expectations. Anonymity exacerbates this violation, as when senders hide their identities, recipients may see them as avoiding accountability (El-Shinnawy & Vinze, 1997; Connolly et al., 1990; Rains, 2007). While most bystanders perceive anonymous hate speech as a clear violation, they mostly remain inactive. This raises the question of whether the aggressor's anonymity influences their decision to intervene.

Anonymous hate speech warrants additional research, since it may not entirely conform to recognized forms of hate speech incidents (Woods & Ruscher, 2021; Wang, 2020). Wang (2020) argues that anonymous hate speech, in comparison to identifiable hate speech, further undermines the perceived credibility of aggressors and the persuasiveness of public discussion in comment forums. Although research confirms anonymity's impact on credibility (Wagenknecht et al., 2016), we do not yet know whether and how the extent of anonymity also affects the intention to intervene of bystanders. Arguably, bystanders may perceive anonymous hate speech commenters as lacking credibility, reducing the perceived threat of their messages. Lastly, the bystander effect states that the more bystanders are present in situations in which a victim needs help, the less likely individuals become to intervene because of diffusion of responsibility (Darley & Latane, 1968). Given that SNS have increased audience sizes compared to offline contexts, the bystander effect may be even more pronounced, underscoring the need for further research.

Anonymity can be understood from two perspectives: the perception of oneself as being anonymous to others (self-anonymity) or the perception of other people being anonymous to oneself (other-anonymity) (Spears and Lea, 1994). This study will be conducted from the perspective of other-anonymity, focusing on the perception of anonymity when reacting to communication from an anonymous source. While considerable research has been conducted

from the perspective of self-anonymity, by for instance studying how anonymous bystanders intervene in online hate speech or to what extent the aggressors' anonymity predicts the perpetration of hate speech, we hardly understand whether and how the degree of anonymity influences the way pure bystanders respond to or perceive anonymous online hate speech. The perception of bystanders on the aggressor's anonymity, hence other-anonymity, on this dynamic remains uncertain, necessitating further investigation into the factors that influence the intention to intervene through counterarguing. Since investigating the consequences of hate speech is important to understand its potential role in destructive social dynamics such as the formation of prejudices or polarized attitudes (Pluta et al., 2023), this study aims to address the following research question: How does the degree of anonymity of online hate speech aggressors influence bystanders' intentions to intervene on SNS?

Theoretical Framework

As previously defined, hate speech refers to online aggression that incites hostility or expresses incivility toward individuals or groups based on certain characteristics (Salminen et al., 2020; Schäfer et al., 2022; Siahaan, 2023). Unlike cyberbullying, which targets individuals repeatedly over extended periods, online hate speech typically occurs in isolated incidents (Leonhard et al., 2018; Obermaier, 2022; Woods & Ruscher, 2021). Although cyberbullying and online hate speech are distinct concepts, we can draw on the literature from cyberbullying due to its similar predictors and consequences (Fulantelli et al., 2022). Hate speech encompasses a wide range of communication forms, such as blasphemy, defamation, provoking, inciting, mocking, publishing false information (Siahaan, 2023), online harassment, and trolling (Álvarez-Benjumea & Winter, 2018).

Drivers of online hate

One explanation for the prevalence of online hate speech is the online disinhibition effect, which describes the phenomenon in which individuals express themselves more freely or aggressively online, often due to perceived anonymity or pseudonymity and lack of real-world consequences (Suler, 2004). This phenomenon can work in two ways: benign disinhibition, where individuals share personal emotions or fears, and toxic disinhibition, which leads to negative behaviours such as hate speech (Suler, 2004). The deindividuation theory explains this further, suggesting that the anonymity provided by SNS reduces individuals' sense of accountability. As a result, their internal restraints and their concerns about negative social evaluations reduce, resulting in behaviors that may not occur in direct interpersonal contacts but would with online contacts (Davidson et al., 2020; Nickerson, 2023; Vilanova et al., 2017).

Moreover, Pabian & Vandebosch (2023) suggest that individuals displaying specific dark personality traits are more likely to experience online moral disengagement, which subsequently increases the likelihood of them perpetrating online hate speech. Online moral disengagement is defined as cognitive techniques that bystanders and aggressors can employ to justify their damaging or immoral (in)actions online. This disengagement mediates the relationship between dark personality traits, such as psychopathy and sadism, associated with the Dark Tetrad, and the increased likelihood of online aggression perpetration (Pabian & Vandebosch, 2023).

Perpetrator–victim–bystander triad

Bystanders could play a critical role in either reinforcing or mitigating the effects of online hate speech. Salmivalli et al. (1996) identified four roles that bystanders can attain: direct assistants of the bully, passive reinforcers, inactive outsiders who remain silent, and active defenders of the victim. Research shows that while most individuals disagree with bullying or hate speech, their actions often fail to align with their attitudes (Salmivalli et al., 1996). In fact, although most students in the class do not engage in bullying, they may indirectly enable its initiation and continuation (Salmivalli et al., 1996). Consequently, bystanders could apply online moral disengagement strategies to justify their inaction. Bystanders could diffuse the responsibility to others, minimize the consequences by saying it is just online hate, or displace the responsibility to others, which are all online moral justification strategies (Bandura, 1999).

Social sanctioning

Bystanders, irrespective of their behavior, represent the social consensus; their actions or silence can reinforce or deconstruct social norms (Craig et al., 2000). This is explained by social sanctioning, which refers to the ways in which society enforces social norms and

regulates behavior through approval or disapproval (Claridge, 2020). Álvarez-Benjumea & Winter (2018) argue that members learn from other members what behavior is approved, and which behavior is expected of people in certain situations. Norms are established when any deviation from the standard behavior requires a sanction. Sanctions serve to uphold social standards by discouraging counter-normative behaviors. These are often enforced by bystanders, whose disapproval helps groups maintain social control (Rashidi et al., 2020). Contrarily, sanctions can be positively valanced when a member of a group is rewarded for their conformity to group norms. Publicly reinforcing social norms by means of counterspeech is expected to be the most effective intervention method in countering online hate speech (Obermaier, 2022), as it shows clear social norms in comment sections on how to interact with others. Additionally, indirect interventions, like flagging or reporting, are suggested to be ineffective in battling online hate speech at large (Mathew et al., 2018) because they lack a direct disapproval from other people in the community (Rashidi et al., 2020). This is the same for inaction, which may lead hate speech aggressors to interpret silence as consent for their message, ultimately continuing a hateful tone in comment sections (Matthew et al., 2018; Rashidi et al., 2020).

Various factors influence whether bystanders intervene or remain passive when facing online hate speech. Belonging to the same social group or having a close relationship with the victim increases the likelihood of intervention (Liebst et al., 2019; Everett et al., 2015). In addition, personal characteristics such as previous experiences with bullying, age, or heightened empathy also influence intention to intervene (Van Cleemput et al., 2014). Moreover, bystanders are more likely to help a female victim than a male one (Walker & Jeske, 2016). Additionally, contextual factors, like the number of bystanders present, can deter intervention through diffusion of responsibility, particularly in online spaces with large audiences (Latané & Darley, 1970; Darley et al., 2009).

The role of anonymity

Anonymity can be defined as the degree to which a communicator perceives the message source as unknown and unspecified (Cho et al., 2012). Previously, the definition of anonymity was considered a binary definition, but this study will adopt more levels of perceived anonymity, suggesting it is more of a continuum than a dichotomous variable (Hite et al., 2014). Eklund et al. (2021) explain that some platforms provide full anonymity (e.g., some chat rooms), pseudonymity (e.g., Reddit usernames), partial identity disclosure (e.g., dating app profiles), or full identifiability (e.g., professional networking sites). Pseudonymity is an alternate identity that could be interpreted as factual or fictitious. While fictitious pseudonyms, like those used in the majority of online chat rooms, are viewed as fictitious by the audience, factual pseudonyms, like those used in aliases, make it impossible for the recipients of a message to detect whether the apparent source is the actual one (Rains & Scott, 2007).

The bystander intervention model (BIM) states that to intervene in an emergency, bystanders need to (1) notice a critical situation, (2) be aware of the fact that the situation is an emergency (to a certain degree), (3) consider themselves personally responsible to intervene, (4) reflect on how to help, and (5) decide to intervene and to implement that decision (Latané & Darley, 1970). People are more likely to interpret situations involving a higher level of harm or danger as emergencies (Reynolds et al., 2023). Anonymous hate speech messages may be seen as less trustworthy, credible (Young et al., 2018), or persuasive (Wagenknecht et al., 2016), potentially reducing their impact on bystanders. Bystanders might infer that when aggressors choose to remain anonymous, they become unaccountable for their message, and the potential harm to victims may also seem less severe, resulting in inaction.

Direct intervention

Bystanders may derogate the aggressor, the message, and the communication platforms when they encounter anonymous content, which could possibly influence their intention to intervene as well (Rains & Scott, 2007). Evaluations of the source of a message impact how messages are perceived and influence bystanders' willingness to intervene in online hate speech incidents. Bonalumi et al. (2018) suggest that receivers can track signals to stabilize communication, and communication sources signal the extent to which they are willing to take responsibility for the contents of their message. By remaining anonymous, aggressors show that they are unwilling to face the consequences of their behaviors to their reputation. Sources can help receivers judge how trustworthy or reliable the message is. Sources can help receivers judge how trustworthy or reliable the message is. When a source is evaluated as uncredible, this also influences how the message is perceived, resulting in the message being perceived as less persuasive (Harmon & Coney, 1982). The choice to remain anonymous can be interpreted by bystanders as avoidance of accountability, potentially signalling to bystanders that direct intervention would be ineffective in intervening in the hate.

Moreover, like the identifiable victim effect, which posits that bystanders may find it harder to empathize with anonymous victims (Jenni & Loewenstein, 1997), anonymity can create psychological distance, which reduces empathy and connection. When *aggressors* lack identifiable characteristics, bystanders may perceive them as less real or human, which can weaken their sense of responsibility to intervene (Lesner & Rasmussen, 2014). In the absence of recognizable, human-like cues, anonymous *aggressors* are likely to feel less human to bystanders, not evoking connection, possibly lowering their intentions to engage in direct intervention.

H1: Aggressor anonymity negatively impacts intention to directly intervene.

In line with this, social presence may explain why anonymity could alter the way arguments are perceived regarding message quality, credibility, and eventually persuasiveness (Wagenknecht et al., 2016). Social presence refers to the extent to which someone feels the presence, closeness, and personality of another person in a digital or virtual environment (Weidlich et al., 2018). The focus is on the sense that there is a "real" person on the other side, even though the interaction occurs through text messages, video chats, or social media (Caroux, 2022). The presence of social cues like a name or personal information increases perceived social presence, which in turn increases feelings of warmth and trust (Wagenknecht et al., 2016). Consequently, social presence can decrease psychological distance between the aggressor and the bystander (Rim et al., 2014), which can increase a bystander's intention to directly intervene. Social Information Processing theory (SIP) suggests that dialogues with fewer verbal and nonverbal cues, which could be interpreted as lower social presence, may take longer to achieve the same levels of intimacy as face-to-face dialogues (Walther, 2015). Adding individual characteristics, like profile photos, helps to lessen anonymity and increase perceived social presence, which enhances credibility.

H2: Aggressor anonymity negatively impacts intention to directly intervene through diminished perceived social presence.

Young et al. (2021) suggest that high perceived anonymity in one-way communication (which is mostly the case with online hate speech), as opposed to two-way communication, negatively impacts credibility in dialogues, whereas in two-way communication there was no difference between identifiable and anonymous sources. Bystanders may believe that online hate speech aggressors are unwilling to take responsibility for their arguments when they choose to remain anonymous (El-Shinnawy & Vinze, 1997) and may question the aggressors' competence and intentions (Connolly et al., 1990). Bystanders might believe their

interventions are less likely to have an impact, and their intention to intervene might not occur.

H3: Aggressor anonymity negatively impacts intention to directly intervene through diminished perceived credibility.

Source credibility theory (Hovland & Weiss, 1951) and its application in online environments (Metzger et al., 2010) posit that the effectiveness of a message depends on the perceived credibility of its source. When an aggressor is anonymous, it lowers their social presence, which provides bystanders limited cues to assess their credibility (Harmon & Coney, 1982). This emphasizes how a lack of identification cues results in a feeling of ambiguity and could result in a lowered perception of social presence. However, receivers of online communication remain motivated to reduce uncertainty about the source they are communicating with and tend to make attributions about the identity of message senders regardless (Wang, 2020). SIP theory suggests that when there is little online information about a person, assessments tend to be more negative and impersonal (Rains, 2007; Rutter, 1987; Walther, 2015). Therefore, perceived anonymity of a hate speech aggressor inhibits social presence, which may reduce bystanders' intention to directly intervene due to a lack of perceived credibility.

H4: Aggressor anonymity negatively impacts intention to directly intervene through diminished perceived social presence and diminished credibility, respectively.

Indirect intervention

Only a small proportion of bystanders directly intervened in instances of cyberbullying, with the majority opting for indirect intervention (i.e., flagging as inappropriate or reporting to the platform) or remaining inactive completely (Dillon & Bushman, 2015). Indirect interventions are approaches that, in comparison to direct interventions, do not directly address a victim or an aggressor and can be private, by blocking aggressors or flagging

hateful content, but can also be public by disliking comments or using likes to support previously posted counterarguments contradicting the hate speech (Citron & Norton, 2011).

Bystanders could interpret anonymous aggressors as avoiding accountability, potentially signalling to bystanders that direct intervention would be ineffective in intervening in the hate speech. However, Jeyagobi et al. (2022) argue that when confronted with anonymous aggressors, bystanders may still feel a moral obligation to intervene without being motivated to use much effort, therefore choosing to use indirect interventions. These indirect interventions are tempting in online environments, especially since online platforms provide convenient, anonymous, and fast methods to report aggressors indirectly (Karasavva & Mikami, 2024). Furthermore, indirect interventions were most chosen for bystanders who did not personally know the victim or had no sufficient knowledge on how to respond (Gahagan et al., 2015). Therefore, while bystanders may be less inclined to directly intervene when the aggressor is anonymous, bystanders may still be inclined to intervene indirectly.

H5: Aggressor anonymity positively impacts bystanders' intention to indirectly intervene.

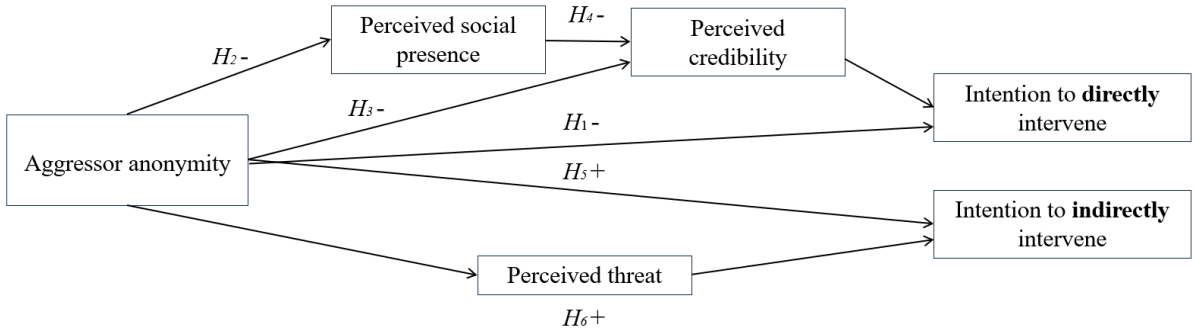
Bystanders may become uncertain about the aggressor's motives when the aggressor of hate speech online is anonymous (Chen et al., 2020). Because anonymity frequently lowers a person's accountability in online settings, spectators may see anonymous aggressors as more inclined to act without restraint or repercussion (Mathew et al., 2019). Bystanders' responses to hate speech could be affected by their perception of threat, even when it is not an imminent danger. Instead of confronting the aggressor directly, which could put them in danger or cause conflict, they might decide to report or flag the content instead, which are safer and less confrontational options. According to Van Cleemput et al. (2014), 31.8% of bystanders named fear of retaliation as a moral disengagement strategy for why they remained inactive when being exposed to cyberbullying. However, bystanders who

intend to mitigate the threat without directly confronting the aggressor may find that indirect intervention methods, such reporting or flagging the hate speech, are an appropriate choice.

H6: Aggressor anonymity positively impacts bystanders' intention to indirectly intervene through a heightened perceived threat.

Figure 1

Conceptual model



Method

To assess the hypotheses, an online experiment was conducted with a between-subject design. An experimental approach was chosen to examine causal relationships between the degree of aggressor anonymity and bystanders' intention to intervene, because this allows for controlled manipulation of variables and ensures internal validity (Bryman, 2016). The independent variable, the degree of anonymity, was studied by manipulating the degree of anonymity of the hate speech aggressor, resulting in three conditions: an identifiable aggressor, a pseudonymous aggressor, and an anonymous aggressor. The platform chosen to imitate in the experiment is X, because of its heightened prevalence of online hate speech (Hickey et al., 2023). The dependent variables in this study were intention to directly intervene and intention to indirectly intervene of bystanders, with perceived threat, perceived credibility, and social presence acting as the mediating variables.

Participants

Using the G-Power 3.1.9.7 program, a minimum sample size of n = 108 was calculated for an effect size of 0.10, alpha = 0.05, and a power of 0.9. In social science research, a power of 0.80 or higher is commonly accepted (Lakens, 2022). In total, 211 Dutch-speaking participants were recruited for the study, which were sampled by means of snowball sampling. However, 32 participants had not finished the experiment completely: 14 participants had not finished a single construct and were therefore excluded. Participants (n = 18) who had at least finished the first two questionnaires measuring both the dependent variables were included for the study. The total sample size consisted of 179 participants, from which 30.4% were males and 67.7% females, with a mean of 36.7 years (SD = 15.3). Of all participants, 68.3% possessed a Bachelor of Applied Science (HBO) degree or above. Except for 6 participants, the sample consisted of people who at least sometimes read

comment sections, from which 51.6% of the participants regularly or more frequently read comment sections.

Pretest

To determine which comment evoked the highest perceived threat while also being considered as realistically occurring in real-life comment sections, a pretest was conducted in which five different comments were shown to seven participants who would not be participating in the study. The hate speech comment in the final experiment was chosen based on the highest perceived threat score ($M = 5.27$, $SD = 1.07$) and the perceived realism score ($M = 5.80$, $SD = 1.29$) and is shown in Figure 2. See Appendix A for the total descriptives of all five comments in the pretest. Additionally, the first manipulation check was measured in the pretest by assessing the perceived anonymity of the three conditions, ensuring there are differences in perceived anonymity between the conditions. Due to a small sample size of seven respondents, which prevents statistical significance testing, the author relied on face validity to evaluate the differences in perceived anonymity across the three conditions.

Figure 2

Stimulus material

Note. Translation: "Body positivity is just an excuse for disgusting pigs to avoid having to change themselves. You're not 'beautiful', you're just fat and lazy!"

The participants in the identifiable condition (Figure 3) were exposed to a hate speech aggressor with a profile picture and an identifiable name including a Dutch area code. The profile picture and the name were chosen by searching for the most average Dutch man and the most frequently occurring Dutch male name. The pseudonymous condition (Figure 4) used an aggressor with a cartooned version of the profile picture used in the identifiable condition as well as a username that was a pseudonym of the real name used in the identifiable condition. Even the smallest identity cues, such as a username or a profile picture, can significantly influence how individuals evaluate other people online (Tanis & Postmes, 2003). Users rely on these limited cues to form impressions and make attributions about communication partners (Spears & Lea, 1994). Aiming to resolve this bias, it was decided to make a cartooned version of the profile picture used in the identifiable condition to minimize impression differences. In Figure 5, the anonymous aggressor did not have a profile picture, "Anonymous" as a name, and an unidentifiable username.

Figure 3

Comment of identifiable aggressor

Figure 4

Comment of pseudonymous aggressor

Figure 5

Comment of anonymous aggressor

Note. The profile picture was collected from the open database of Canva and turned into a cartoon using Picsart.

Procedure

In the final Qualtrics' web experiment, participants had to give their informed consent to participate in the study first. The informed consent included a disclaimer that this experiment used a fictional comment and news article generated by ChatGPT and would take approximately ten minutes to finish. It was chosen to disclose the fact that the stimulus material was fictional in advance because of the Responsibility pillar of the Code of Conduct.

After informed consent, participants were randomly allocated to one of three conditions. The experiment began with a tweet from the X profile of a Dutch newspaper about body positivity, attached in Appendix B, including a time slot to ensure respondents carefully assessed the stimulus material. After the tweet followed a retweet, with depending on the condition they were assigned to, a comment from an anonymous aggressor, a pseudonymous aggressor, or a comment from an identifiable aggressor was shown (see Figure 3, 4 and 5). After the stimulus material, intention to directly and indirectly intervene was measured first, after which the mediators perceived threat, social presence, and credibility were measured in questionnaires with multi-item constructs. Between the measurement of the dependent variables and the mediating variables, the stimulus materials were shown again to refresh the memory of the participants. Following the measurement of

the mediating variables, the control variables of perceived anonymity and identification with body positivity were measured. The experiment ended with gathering demographic information, such as gender, age, educational level, and frequency of reading comments. Lastly, the questionnaire was closed with a debriefing stating that all materials used in the experiment were fictional and their data will be used anonymously.

Measures

To improve the accuracy of comprehension of the measures for a Dutch-speaking sample, backtranslation was employed to successfully translate the English constructs to Dutch (Brislin, 1970). First, the scales were translated from English to Dutch by the author, after which ChatGPT was employed to translate the Dutch version back into English. Considering the differences in these two translations, the items were reviewed and refined to be culturally appropriate while retaining the original meaning. Direct and indirect intervention intentions were rated on 5-point Likert scales ranging from 1 (not likely) to 5 (very likely), whereas the other measures were rated on 7-point Likert scales, ranging from 1 (strongly disagree) to 7 (strongly agree). According to Boateng et al. (2018), using multiple different scales in questionnaires could cause confusion amongst participants. Therefore, a 7-point scale or a 5-point scale was chosen to employ for the mediating and control variables, because these scales offer participants the possibility to provide their most accurate and representative judgment compared to shorter-ranging scales or scales without a “middle” point (Finstad, 2010). See Appendix C for the total constructs and questionnaires.

Direct intention to intervene is defined as all direct methods taken by bystanders to respond to hate speech uttered towards victims. As is shown in the theoretical framework, there are multiple options to employ for bystanders to intervene in online hate speech contexts. To ensure all possible direct interventions were included in the scale, the interventions from Obermaier (2022) and Jia and Schumann (2023) were combined to

construct the scale. This includes items around constructive counter speech (e.g., using facts) like “I would refute the statements with facts,” destructive counter speech (e.g., using hate speech in return), but also private measures like confronting the aggressor or supporting the victim in a private message. The reliability of this scale was good, Cronbach’s α = .91 (*M* = 1.89, *SD* = 0.95).

Indirect intention to intervene is defined as interventions that, in comparison to direct interventions, do not directly address a victim or an aggressor. Indirect intervention can be private by blocking aggressors or flagging hateful content, but it can also be public by disliking comments or using likes to support previously posted counterarguments contradicting the hate speech aggressor (Citron & Norton, 2011). To ensure all possible indirect interventions were included in the scale, the interventions from Obermaier (2022), Citron & Norton (2011), and Jia and Schumann (2023) were combined to construct the scale. This included four items like “I would report the hate speech comment to the X platform.” The reliability of this scale was good, Cronbach’s α = .81 (*M* = 2.71, *SD* = 1.46).

Perceived credibility is defined as the extent to which bystanders subjectively experience the source of a message as believable, trustworthy, and accountable. This is measured using the perceived credibility scale (Chesney & Su, 2010), from which the items adjusted to the online source context from Metzger et al. (2010) were used. The scale consisted of 7 items like “I perceive the hate speech aggressor to be trustworthy.” The reliability of this scale was good, Cronbach’s α = .85 (*M* = 3.39, *SD* = 1.12).

Perceived threat is defined as the extent to which bystanders consider the hate speech commenters as threatening. In the context of online hate speech, measuring participants’ risk of intervening, fear of negative consequences, and both psychological and actual threat were used to construct this scale. The scale included 8 items like “I feel that intervening directly

could make me a target for retaliation." The reliability of this scale was good, Cronbach's α = .89 (*M* = 5.25, *SD* = 1.07).

Perceived social presence is defined as the extent to which someone feels the presence, closeness, and personality of another person in a digital or virtual environment (Weidlich et al., 2018). The focus is on the sense that there is a "real" person on the other side, even though the interaction occurs through text messages, video chats, or SNS (Caroux, 2022). This variable was measured using the social presence scale (Weidlich et al., 2018), which consisted of 6 items like "The aggressor felt so 'real' that I almost believed that we were not virtual at all." The scale was adjusted to match the hate speech context, and the reliability of this scale was good, Cronbach's α = .76 (*M* = 3.57, *SD* = 1.03).

Perceived social similarity is defined as the extent to which bystanders identify with the victims of the hate speech to control for group membership. The participants were asked to what extent they identified with the body positivity movement, using the group identification scale altered to fit the body positivity context (Doosje et al., 1995). The scale includes 5 items like "I identify with the body-positivity movement." The reliability of this scale was good, Cronbach's α = .86 (*M* = 3.24, *SD* = 1.29).

Perceived anonymity is defined as the extent to which a bystander considers an aggressor of hate speech as identifiable or anonymous, and the scale is adapted from the Perceived Anonymity Measurement Instrument from (Hite et al., 2014). To ensure the experimental manipulation of anonymity levels (identifiable, pseudonymous, and anonymous) was effective, a manipulation check was conducted. Participants were asked to rate the perceived anonymity of the aggressor using a 7-point Likert scale ranging from 1 (not at all anonymous) to 7 (completely anonymous), including 5 items like "I believe that the personal

identity of the hate speech aggressor is unknown to me." The reliability of the perceived anonymity scale was good, Cronbach's α = .88 (*M* = 4.21, *SD* = 1.38).

Manipulation check

To check whether the manipulation of the conditions differed significantly enough, a one-way ANOVA was conducted to test whether a main effect exists on aggressor anonymity and perceived anonymity. Due to non-normal distribution in the pseudonymous condition (*z*-score skewness / *z*-score kurtosis = -0.716 and 0.377), a Kruskal-Wallis test was performed. The assumption of homogeneity was met (*VR* = 1.045). The Kruskal-Wallis test revealed a significant effect of aggressor anonymity on perceived anonymity scores ($\chi 2$ = 16.4, df = 2, *p* = <.001). A Dwass-Steel-Critchlow-Flinger analysis revealed that the identifiable condition differed significantly from the anonymous condition (*W* = 5.51, p = <.001, d = -0.888). No significant differences were found for the pseudonymous (identifiable vs. pseudonymous: *W* = 3.78, *p* = .021, *d* = -0.524; pseudonymous vs. anonymous: *W* = 2.45, *p* = .194, *d* = -0.364), indicating that this manipulation did not effectively represent the intended middle level of perceived anonymity. Including data from an ineffective manipulation could potentially lead to misleading results and incorrect conclusions about the effects of anonymity on the dependent variables. Therefore, to maintain the integrity of the research and ensure that the analysis focuses on clearly distinct levels of anonymity, the pseudonymous condition was excluded from further analysis, allowing for a more accurate assessment of the effects of anonymity between the anonymous and identifiable conditions.

Results

After excluding the pseudonymous condition from the experiment, a 2 x 1 between-subject analysis was conducted using a MANOVA to study the main effects. To assess the hypotheses, mediation analyses model 4 and model 6 from Hayes (2022) were employed. In Table 1, descriptive statistics of the dependent variables (intention to directly or indirectly intervene), mediating variables (perceived social presence, credibility, and threat), and the control variables (age and identification) are shown per condition (identifiable and anonymous). The standard descriptives show that the participants had low intentions to directly intervene and intended to intervene more indirectly, regardless of the assigned condition. Female participants had higher direct intervention intentions (M = 1.92, SD = 0.891) than male participants (M = 1.58, SD = 0.760). Additionally, female participants also had higher indirect intervention intentions (M = 2.88, SD = 1.533) than male participants (M = 2.23, SD = 1.335).

Table 1

Descriptive statistics on a 5-point or 7-point scale

	Aggressor Anonymity		
	Identifiable	Anonymous	*N*
Intention to indirectly intervene	2.47 (*SD* = 1.365)	2.94 (*SD* = 1.527)	115
Intention to directly intervene	1.87 (*SD* = 1.003)	1.90 (*SD* = 0.908)	115
Perceived threat	5.03 (*SD* = 0.943)	5.45 (*SD* = 0.981)	101
Perceived anonymity	3.61 (*SD* = 1.280)	4.73 (*SD* = 1.251)	101
Perceived credibility	3.60 (*SD* = 1.189)	3.20 (*SD* = 1.036)	101
Perceived social presence	3.73 (*SD* = 1.104)	3.43 (*SD* = 0.954)	101
Perceived similarity (in-group/out-group)	3.17 (*SD* = 1.323)	3.29 (*SD* = 1.275)	101
Age	43 (*SD* = 16.6)	36.7 (*SD* = 14.9)	101

Table 2 shows the standard descriptives of all individual intervention intentions to provide a more thorough overview. Participants who intended to directly intervene in hate speech, either from identifiable or anonymous aggressors, mostly intended to employ approaches that focused on supporting the victim (items 6, 7, and 12). Additionally, constructive counterarguing methods were the most frequently chosen approaches (items 1, 2, 3, and 4), whereas destructive counterspeech methods were the least chosen intended approaches (items 8, 9, 10).

Table 2

Descriptive statistics on a 5-point scale

	Aggressor Anonymity	
	Identifiable	Anonymous
1. Direct interventions		
1. Condemning hate speech with factual argumentation	2.00 (*SD* = 1.680)	2.09 (*SD* = 1.430)
2. Responding to hate speech with a question	2.09 (*SD* = 1.672)	1.95 (*SD* = 1.303)
3. Sanctioning the communication by reminding to politeness	2.18 (*SD* = 1.638)	2.14 (*SD* = 1.594)
4. Educating the aggressor	1.81 (*SD* = 1.274)	2.02 (*SD* = 1.445)
5. Trying to persuade the aggressor to change opinion	1.74 (*SD* = 1.247)	1.81 (*SD* = 1.263)
6. Comment to show support and empathy for victim	2.33 (*SD* = 1.694)	2.34 (*SD* = 1.540)
7. Comment to gather more support for victim	2.23 (*SD* = 1.581)	2.21 (*SD* = 1.484)
8. Responding to content of hate speech with similar offensive words	1.47 (*SD* = 0.947)	1.50 (*SD* = 1.047)
9. Responding to the aggressor of hate speech with hate speech	1.35 (*SD* = 0.641)	1.48 (*SD* = 0.819)
10. Insulting the hate speech aggressor	1.37 (*SD* = 0.794)	1.48 (*SD* = 0.883)
11. Confront the speech aggressor in a direct message	1.51 (*SD* = 0.984)	1.69 (*SD* = 1.217)
12. Support the victim in a direct message	2.42 (*SD* = 1.690)	2.12 (*SD* = 1.579)

Table 3 shows that participants intended to indirectly intervene more with anonymous aggressors than with identifiable aggressors. Participants chose intervention methods that did not require reporting the profile or comment to the X platform.

Table 3

Descriptive statistics on a 5-point scale

		Aggressor Anonymity	
		Identifiable	Anonymous
2.	***Indirect interventions***		
1.	Reporting profile of the aggressor to the X platform	1.82 (*SD* = 1.42)	2.47 (*SD* = 1.55)
2.	Reporting the comment to the X platform	2.30 (*SD* = 1.58)	2.76 (*SD* = 1.70)
3.	Liking other comment(s) that contradict the hate speech in the comment section	3.04 (*SD* = 2.13)	3.60 (*SD* = 2.17)
4.	Reacting to comment with non-verbal cues (dislikes, sad emojis)	2.74 (*SD* = 1.87)	2.93 (*SD* = 1.97)

To examine the relationships between the variables in this study, a correlation analysis was conducted. The Shapiro-Wilk test revealed that in the identifiable condition, intention to directly intervene (W = 0.813, p = <.001), intention to directly intervene (W = 0.902, p = <.001), identification (W = 0.927, p = .006), and age (W = 0.871, p = <.001), followed a non-normal distribution. In the anonymous condition, the Shapiro-Wilk test revealed that intention to directly intervene (W = 0.872, p = <.001), intention to indirectly intervene (W = 0.935, p = .004), perceived threat (W = 0.927, p = <.019), identification (W = 0.942, p = .011), and age (W = 0.900, p = <.001), followed a non-normal distribution. This means the assumption of normality is violated, and to resolve these normality issues in the correlation matrix, Spearman's analyses were conducted. The results are summarized in Table 4.

The correlation analysis revealed a positive correlation between direct and indirect intervention intentions ($\rho = 0.601$, $p < .001$), suggesting that bystanders who are likely to engage in one form of intervention could be similarly inclined towards the other type of intervention. Additionally, perceived threat was significantly associated with both direct ($\rho = 0.201$, $p < .044$) and indirect intervention intentions ($\rho = 0.372$, $p = .001$), indicating that heightened perceptions of threat may evoke bystanders' intention to intervene more. Social presence was weakly but significantly positively correlated with perceived threat ($\rho = 0.219$, $p = .028$) but not with both types of intervention intentions.

Perceived credibility showed a strong positive correlation with perceived social presence ($\rho = 0.441$, $p < .001$) but did not significantly correlate with intervention intentions. Perceived anonymity was negatively correlated with perceived social presence ($\rho = -0.276$, $p = .005$) and perceived credibility ($\rho = -0.206$, $p = .039$), indicating that as anonymity increased, both social presence and credibility decreased. However, perceived anonymity did not significantly correlate with either direct or indirect intervention intentions.

Perceived identification was positively correlated with perceived threat ($\rho = 0.359$, $p = < .001$), indirect intervention intentions ($\rho = 0.342$, $p < .001$), and direct intervention intentions ($\rho = 0.254$, $p = .010$), suggesting that identifying with the aggressor may increase perceived threat and intention to intervene both directly and indirectly.

Finally, age was weakly but significantly correlated with perceived credibility ($\rho = 0.198$, $p = .047$), but it did not significantly correlate with intervention intentions or other key variables.

Table 4

Correlation matrix

		Intention to indirectly intervene	Intention to directly intervene	Perceived threat	Perceived social presence	Perceived credibility	Perceived social anonymity	Identification	Age
Intention to indirectly intervene	Spearman's rho	–							
	df	–							
	p-value	–							
Intention to directly intervene	Spearman's rho	0.601***	–						
	df	113	–						
	p-value	< .001	–						
Perceived threat	Spearman's rho	0.201*	0.372***	–					
	df	113	99	–					
	p-value	<.044	<.001	–					
Perceived social presence	Spearman's rho	0.119	0.047	0.219*	–				
	df	99	99	99	–				
	p-value	.236	.639	.028	–				
Perceived credibility	Spearman's rho	0.060	-0.085	0.081	0.441***	–			
	df	99	99	99	99	–			
	p-value	.236	0.399	.420	<.001	–			
Perceived anonymity	Spearman's rho	-0.073	-0.001	0.030	-0.276**	-0.206*	–		
	df	99	99	99	99	99	–		
	p-value	.469	.765	.005	.005	.039	–		
Identification	Spearman's rho	0.254*	0.324***	0.359***	0.217*	-0.018	-0.001	–	
	df	99	99	99	99	99	99	–	
	p-value	.558	<.001	<.001	.029	.857	.989	–	
Age	Spearman's rho	0.059	-0.049	-0.149	-0.119	0.198*	0.080	-0.068	–
	df	99	99	99	99	99	99	99	–
	p-value	.558	.625	0.136	.046	.047	.426	0.502	–

Note. * p <.05, ** p <.01, *** p. <.00

A MANOVA was conducted to examine the main effect of aggressor anonymity on bystanders' intention to directly intervene and indirectly intervene. The Shapiro-Wilk test confirmed violations of the assumption of normality ($W = 0.910$, $p < .001$). Therefore, a more robust test, Pillai's Trace, was employed. The MANOVA revealed that aggressor anonymity did not significantly affect the combined dependent variables of direct and indirect intervention intentions (Pillai's Trace = 0.0316, $F(2,112) = 1.83$, $p = .165$). Additionally, Wilks' Lambda provided similar results (Wilks' Lambda = 0.968, $F(2,112) = 1.83$, $p = .165$). This suggests that anonymity does not have a statistically significant overall impact when compared to both types of intervention. More specifically, the effect of aggressor anonymity on both the dependent variables was examined individually from each other. For indirect intervention intention, the effect was not statistically significant ($F(1,113) = 2.97$, $p = .087$, $\eta p2 = 0.03$), and the effect for direct intervention intention could be considered negligible ($F(1,113) = 0.02$, $p = .894$, $\eta p2 = 0.01$). Overall, these results suggest that aggressor anonymity does not significantly impact either of the two intentions to intervention methods.

Intention to directly intervene

To examine the first four hypotheses examining effects of aggressor anonymity on intention to directly intervene, a sequential mediation analysis was performed using Model 6 from Hayes (2022), including perceived similarity as a covariate. The mediators used in the analysis were perceived social presence and perceived credibility. Because of normality problems within the conditions and variables, the analysis was bootstrapped 3000 times BC.

Hypothesis 1 stated that aggressor anonymity would negatively impact bystanders' intention to directly intervene. The results showed that the direct effect of aggressor anonymity on intention to directly intervene, after controlling for perceived similarity, was not significant ($b = 0.102$, $SE = 0.178$, $p = .567$), suggesting aggressor anonymity does not

directly influence intervention intentions. This is in line with the results of the main effects from the MANOVA analysis, finding no support for Hypothesis 1.

Hypothesis 2 stated that the relationship between aggressor anonymity and direct intervention would be mediated by diminished social presence. The analysis revealed that after controlling for perceived similarity, the path from aggressor anonymity to social presence was not significant ($b = 0.337$, $SE = 0.200$, $p = .092$), nor was the path from social presence to direct intervention ($b = 0.017$, $SE = 0.108$, $p = .873$). Consequently, the indirect effect through social presence was not significant ($b = 0.006$, $SE = 0.043$, $p = .892$). These findings indicate that social presence does not mediate the relationship between aggressor anonymity and intention to directly intervene, finding no support for Hypothesis 2.

Hypothesis 3 stated that the relationship between aggressor anonymity and direct intervention would be mediated by diminished perceived credibility. Additionally, Hypothesis 4 suggested that aggressor anonymity would indirectly impact intention to directly intervene through diminished social presence and reduced credibility, respectively. The results showed that the path from aggressor anonymity to credibility was approaching significance ($b = 0.395$, $SE = 0.219$, $p = .071$), and the path from credibility to direct intervention was not significant ($b = 0.043$, $SE = 0.079$, $p = .587$). The combined indirect effect from aggressor anonymity, through social presence, to credibility on intention to directly intervene was not significant ($b = -0.002$, $SE = 0.039$, $p = .953$). These findings indicate that neither credibility alone nor the combination of social presence and credibility mediates the relationship between aggressor anonymity and intention to directly intervene, finding no support for Hypothesis 3 and Hypothesis 4.

The total effect of aggressor anonymity on intention to directly intervene, combining both direct and indirect effects, was also not significant ($b = 0.102$, $SE = 0.178$, $p = .567$).

These results suggest that aggressor anonymity does not significantly impact bystanders' intention to directly intervene, either directly or through the mediators of social presence and credibility.

Intention to indirectly intervene

To examine Hypotheses 5 and 6, a mediation analysis was performed using Model 4 from Hayes (2022), including perceived similarity as a covariate. Hypothesis 5 stated that aggressor anonymity positively impacts pure bystanders' intention to indirectly intervene, and Hypothesis 6 suggested that a heightened perceived threat mediated this relationship. Because of normality problems within the conditions and variables, the analysis was bootstrapped 3000 times BC. The direct effect of aggressor anonymity on indirect intervention intentions was not significant (b = 0.254, SE = 0.367, p = 0.359), indicating that anonymity alone does not directly influence bystanders' likelihood of indirect intervention. This is in line with the results from the main effects from the MANOVA analysis, indicating that Hypothesis 5 is not supported.

However, the mediation analysis revealed that aggressor anonymity significantly influenced perceived threat (b = 0.380, SE = 0.176, p = .030), because anonymous aggressors were perceived as more threatening than identifiable ones. In turn, perceived threat significantly predicted intention to indirectly intervene (b = 0.375, SE = 0.169, p = .027), suggesting that higher levels of perceived threat were associated with greater intentions to intervene indirectly. However, the indirect effect of aggressor anonymity on indirect intervention intentions via perceived threat was not significant (b = 0.143, SE = 0.105, p = 0.173). These results partially support H6, as perceived threat significantly predicts indirect intervention intentions, and anonymity indirectly reduces these intentions through perceived threat. However, the indirect effect is not significant. The total effect, combining both direct and indirect effects, was not significant (b = 0.397, SE = 0.286, p = 0.165).

Additionally, perceived similarity did significantly influence both perceived threat (*b* = 0.291, *SE* = 0.071, *p* = <.001) and indirect intervention intentions (*b* = 0.263, *SE* = 0.111, *p* = .017). This suggests that the more participants identified with the targeted group of the hate speech, the more their intervention intentions would indirectly increase by heightened perceptions of threat.

Discussion

Although multiple studies (Obermaier, 2022; Yu, 2024; Zapata et al., 2024) suggest that bystanders can play a crucial role in reducing online hate speech, this experiment, alongside other studies, revealed that bystanders rarely intervened in such incidents. Obermaier (2022) expected that counterarguing from bystanders would be essential in mitigating online hate speech, as it functions as a normative control method that publicly disapproves of hostile behavior and establishes clear behavior rules. Furthermore, bystander intervention was considered more effective than automated detection because it upholds free speech principles while also sanctioning the hate speech (Mathew et al., 2018; Rashidi et al., 2020; Wang et al., 2020), thereby preventing aggressors from interpreting bystanders' silence as consent. However, when bystanders did not belong to the targeted group or have a personal relationship with the victim, which is mostly the case with online hate speech, their intervention intentions were relatively low (Liebst et al., 2019; Everett et al., 2015). This finding necessitates further research into other variables possibly affecting bystanders' responses.

The effect of anonymity on perpetrators or bystanders' own communication behavior (self-anonymity) has been extensively studied, though a noted gap in the literature concerns how bystanders perceive anonymous communicators, hence other-anonymity (Rains & Scott, 2007; Wang, 2020; Woods & Ruscher, 2021). This study aimed to fill in this gap and investigate whether aggressor anonymity (identifiable, pseudonymous, or anonymous) influences bystanders' intentions to directly and indirectly intervene against online hate speech. However, the pseudonymous condition (n = 64) was eliminated from the study due to its excessive overlap with the other two conditions and failure to pass the manipulation check, leaving a 2x1 between-subjects web design (n = 115). Understanding how bystanders

interpret and respond to different levels of anonymity remains critical for addressing the social dynamics of online communication. However, while aggressor anonymity had no significant direct effects on either direct or indirect intention to intervene, several key findings contribute to the existing body of research.

The results showed the existence of a bystander effect where individuals are reluctant to intervene in situations where others are also present (Darley & Latané, 1968). Overall, participants demonstrated low intentions to indirectly intervene and even lower intentions to directly address the aggressor, even in the context of a fictional experiment. This is worrisome, as the intention-behavior gap suggests that low intention rates frequently translate to even lower actual intervention rates (Conner & Norman, 2022). These low intentions to directly intervene, such as through counterspeech or supporting the victim, might be explained by the identifiable victim effect (IVE) (Jenni & Loewenstein, 1997). This phenomenon describes people's tendency to provide more assistance to identifiable, specific victims rather than to a larger, anonymous group. In this experiment, the hate speech targeted all members of the body positivity movement rather than a single identifiable victim, which likely reduced bystanders' empathy and created psychological distance, reducing bystanders' intention to intervene (Zhao et al., 2024). Lastly, the experiment did not involve a victim personally known to the participants, which could have influenced their intervention intentions, as bystanders are more likely to intervene when they have a pre-existing relationship with the victim (Liebst et al., 2019). This absence of direct intervention is consistent with theories of moral disengagement, which hold that bystanders justify their inaction by misrepresenting the consequences for victims or diffusing responsibility (Bandura, 1999).

Aggressor anonymity did not appear to account for the low level of direct interventions observed in this experiment. The results suggest the presence of a floor effect

for direct interventions: when bystanders are confronted with hate speech they generally refrain from intervening, regardless of aggressor anonymity. Moreover, it was hypothesized that perceived credibility and social presence would mediate this relationship between aggressor anonymity and intention to directly intervene, yet these factors were found insignificant. Rains (2007) demonstrated that anonymous sources can be perceived as similarly credible as identifiable sources in the context of online health information. Likewise, Chesney (2010) found that the credibility of a blog was influenced more by the quality of its presentation than by the writer's anonymity. These findings suggest that the impact of anonymity on credibility may be less pronounced in certain contexts, potentially explaining why credibility does not explain intervention intentions in this study.

Furthermore, the character of the SNS messages used in this experiment could explain why social presence was not a significant mediator. The experiment used a single, short social media post with a reading time of eight seconds. According to Weidlich et al. (2018), social presence refers to the extent to which someone feels the presence and closeness of another person in a digital environment. Richer media formats, like video or text-based conversations, might be better for evoking social presence. Short exposure could possibly not provide bystanders sufficient information for experiencing this presence in the context of rapid and detached social media interactions (Lowenthal, 2010), possibly explaining why social presence was consistently low, regardless of aggressor anonymity.

Nevertheless, the real cause(s) of these low direct intention rates are still unknown and need more research. One potential explanation is desensitization to aggressive content (Soral et al., 2017). Desensitization studies show that frequent exposure to hate speech diminishes sensitivity to such content, resulting in lower sympathy for the victims (Soral et al., 2017) or lower empathetic responsiveness from bystanders (Pabian et al., 2016). This loss of empathy, which has been shown as a predictor of helping the victim in cyberbullying contexts (Van

Cleemput et al., 2014), is a crucial characteristic that future studies should focus on to reduce desensitization.

Another phenomenon that could explain why bystanders of online hate speech may refrain from direct intervention is self-censorship (Sweeney, 2003). Bystanders often weigh the social costs of intervening, such as the potential for conflict, leading them to adjust their responses or refrain from action altogether by using a "filter," while being aware of the harmful nature of the content (Warner & Wang, 2019). The preference for anonymity, combined with concerns about personal reputation and the fear of being targeted themselves, encourages bystanders to self-censor their intentions to act (Davidovic et al., 2023). This aligns with the item-level analysis in this study, which revealed that bystanders primarily chose intervention methods that avoided confrontation with the aggressor (like destructive and constructive counterspeech) and mostly chose to intervene by supporting the victim. The heightened perceived risks of engaging in destructive counterspeech, including the increased likelihood of conflict, illustrate the role of self-censorship. However, this self-censorship did not mean bystanders remained completely passive.

Despite low rates of direct intervention, participants were inclined to engage in indirect methods more, such as reporting or flagging offensive content. These reporting mechanisms provide users an active role in shaping the conversational tone (Crawford & Gillespie, 2014). The results revealed that bystanders perceived anonymous aggressors as more threatening than identifiable aggressors, which in turn evoked more indirect intervention intentions, yielding support for H6. This is in line with multiple studies suggesting that the more hate speech situations are perceived as threatening, the higher bystanders' intention or willingness to intervene is (Leonhard et al., 2018; Koehler & Weber, 2018). In addition to these studies, Wachs et al. (2017) suggest that in situations of heightened threat perceptions, the fear of relation is one of the key motivators for bystanders to opt for indirect interventions.

Although indirect interventions are suggested to be less effective than direct interventions in reducing online hate speech, these still represent useful intervention methods for bystanders. While indirect methods may not immediately counteract hate speech or diminish negative outcomes for victims, they can contribute valuable data to automated hate speech detection systems (Azam et al., 2022). With developments in data collection and the rise of more sophisticated AI-driven bots, these systems can increasingly identify and address hate speech in real time. This is becoming even more critical in managing hate speech effectively, with the decreasing availability of human moderators on many platforms (Gillespie, 2019). As a result, these innovations could reduce the reliance on human moderators and foster a less aggressive online environment.

To enhance bystanders' participation in maintaining a constructive online environment, improving digital literacy could be essential. Less digitally literate individuals may not be aware that these indirect intervention mechanisms are available or that they experience uncertainty as to what extent their interventions are employed anonymously (Crawford & Gillespie, 2014). This lack of awareness and uncertainty can significantly hinder intervention efforts, and according to Osama (2023), targeted educational programs centred around improving digital literacy have been found to improve bystander intervention. Consequently, Davidovic et al. (2023) highlight that some bystanders hesitate to intervene indirectly due to uncertainty about the outcomes of their actions. This lack of feedback regarding the effectiveness of their reporting or flagging discouraged some individuals from intervening, and as a result remained inactive (Wong et al., 2016). Future experimental studies could explore whether providing feedback on the results of their reports, by for example, informing users that their flagging or reporting resulted in the removal of a hateful post, increases bystanders' willingness to engage in indirect interventions.

This study had three main limitations that would need consideration. First, the sample was relatively small (n = 179; of which 64 were excluded) and predominantly composed of female participants (30.4% males, 67.7% females). This distribution of gender may have influenced the results, as prior research indicates that female bystanders are generally more likely to intervene than male bystanders, specifically in low- and high-harm situations (Walker & Jeske, 2016). Contrarily, male bystanders tend to intervene more often when the victim is female, especially in high-harm scenarios. Future research could address this limitation to ensure a more balanced sample to increase generalizability.

Second, the pseudonymous condition failed to differ significantly from the identifiable condition, leading to the exclusion of one experimental condition. Although it was attempted to create clear distinctions between the conditions, the overlap observed suggests potential differences in the perception of how anonymous pseudonymous sources are seen. Perceived anonymity, rather than actual anonymity, often drives online behavior (Hite et al., 2014). In this study, perceived anonymity was measured by using items that asked to what extent the participants thought they were able to identify the aggressor. This ability to detect someone's identity could be mediated by digital literacy. As users increase their ability to identify identities from pseudonymous profiles, it could potentially reduce the uniqueness of this condition. Nevertheless, larger samples are needed in future research to study subtle differences in aggressor anonymity perceptions.

Lastly, while participants participated in the experiment in a natural environment (i.e., on their own devices, in their personal environments), a field experiment could have provided more ecologically valid insights. However, conducting such an experiment raises significant ethical concerns, as field experiments can present various harms to individuals who did not consent to participation, undermining autonomy and potentially damaging public trust

(Phillips, 2021). Instead, future research could utilize existing data or interview moderators who actively perform these tasks to find more insights.

Conclusion

Although scholars advocate for bystander intervention as a crucial strategy to mitigate online hate speech (Obermaier, 2022; Yu, 2024; Zapata et al., 2024), relying on bystanders to be social norms setters may be overly idealistic. Most bystanders refrain from direct interventions, especially in the absence of identification with the targeted group or a personal connection to the victim. Therefore, relying on direct intervention as the primary solution to mitigate hate speech is both impractical and ineffective. In contrast, empowering bystanders to engage in effective indirect interventions does seem to be more realistic and effective.

This experiment demonstrated that anonymous aggressors were perceived as more threatening than identifiable aggressors, leading to heightened perceptions of threat and stronger intentions to engage in indirect intervention intentions. Improving bystanders' digital literacy could enhance their ability to indirectly intervene in online hate speech. Educating bystanders about their own anonymity when using reporting mechanisms could reduce their fear of retaliation, which encourages proactivity while responding to the hate speech. Increased awareness of reporting mechanisms, combined with clear feedback on the outcomes of their intervention, can foster more participation in intervening. Furthermore, the data generated through these indirect interventions could play a crucial role in the refinement of automated hate speech detection systems. Considering recent technological advancements, empowering bystanders to intervene indirectly represents a practical, scalable, and realistic solution to fostering safer online environments.

In conclusion, while bystanders can counter online hate speech through direct intervention, they frequently refrain from intervening directly. Relying on bystanders as norm

reinforcers is therefore an ineffective and overly idealistic approach. Instead, focusing on bystanders as frontline detectors of online hate speech represents a more realistic solution. Focusing on indirect actions and enhancing digital literacy can bridge the gap between human and automated efforts to effectively combat online hate speech.

References

Agarwal, P., Hawkins, O., Amaxopoulou, M., Dempsey, N., Sastry, N., & Wood, E. (2021). Hate Speech in Political Discourse: A Case Study of UK MPs on Twitter. *Conference on Hypertext and Social Media*. https://doi.org/10.1145/3465336.3475113

Álvarez-Benjumea, A., & Winter, F. (2018). Normative Change and Culture of Hate: An experiment in online environments. *European Sociological Review*, *34*(3), 223–237. https://doi.org/10.1093/esr/jcy005

Azam, U., Rizwan, H., & Karim, A. (2022). Exploring data augmentation Strategies for hate speech detection in Roman Urdu. *ACL Anthology*. https://aclanthology.org/2022.lrec-1.481

Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review*, *3*(3), 193–209. https://doi.org/10.1207/s15327957pspr0303_3

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best Practices for developing and Validating scales for health, Social, and Behavioral Research: A primer. *Frontiers in Public Health*, *6*. https://doi.org/10.3389/fpubh.2018.00149

Bonalumi, F., Isella, M., & Michael, J. (2018). Cueing implicit commitment. *Review of Philosophy and Psychology*, *10*(4), 669–688. https://doi.org/10.1007/s13164-018-0425-0

Brislin, R. W. (1970). Back-Translation for Cross-Cultural research. *Journal of Cross-Cultural Psychology*, *1*(3), 185–216. https://doi.org/10.1177/135910457000100301

Bryman, A. (2016). *Social research methods*. Oxford University Press.

Castaño-Pulgarín, S. A., Suárez-Betancur, N., Vega, L. M. T., & López, H. M. H. (2021). Internet, social media and online hate speech. Systematic review. *Aggression and Violent Behavior*, *58*, 101608. https://doi.org/10.1016/j.avb.2021.101608

Chen, X., Huang, C., & Cheng, Y. (2020). Identifiability, risk, and information credibility in discussions on Moral/Ethical Violation topics on Chinese social networking sites. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.535605

Chesney, T., & Su, D. K. (2010). The impact of anonymity on weblog credibility. *International Journal of Human-Computer Studies*, *68*(10), 710–718. https://doi.org/10.1016/j.ijhcs.2010.06.001

Cho, D., Kim, S., & Acquisti, A. (2012). Empirical analysis of online anonymity and user behaviors: The impact of real name Policy. *Journal of Methods and Measurement in the Social Science*. https://doi.org/10.1109/hicss.2012.241

Cho, H., Shen, L., & Wilson, K. (2012). Perceived realism. *Communication Research*, *41*(6), 828–851. https://doi.org/10.1177/0093650212450585

Citron, D. K., & Norton, H. L. (2011). Intermediaries and Hate Speech: Fostering Digital Citizenship for our information Age. *Boston University Law Review*, *91*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1764004

Claridge, T. (2020). Social sanctions – overview, meaning, examples, types and importance. *CERN European Organization for Nuclear Research*. https://doi.org/10.5281/zenodo.8016175

Conner, M., & Norman, P. (2022). Understanding the intention-behavior gap: The role of intention strength. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.923464

Connolly, T., Jessup, L. M., & Valacich, J. S. (1990). Effects of anonymity and evaluative tone on idea Generation in Computer-Mediated Groups. *Management Science*, *36*(6), 689–703. https://doi.org/10.1287/mnsc.36.6.689

Correa, D., Silva, L., Mondal, M., Benevenuto, F., & Gummadi, K. (2021). The many shades of anonymity: characterizing anonymous social media content. *Proceedings of the International AAAI Conference on Web and Social Media*, *9*(1), 71–80. https://doi.org/10.1609/icwsm.v9i1.14635

Craig, W. M., Pepler, D., & Atlas, R. (2000). Observations of bullying in the playground and in the classroom. *School Psychology International*, *21*(1), 22–36. https://doi.org/10.1177/0143034300211002

Crawford, K., & Gillespie, T. (2014). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society*, *18*(3), 410–428. https://doi.org/10.1177/1461444814543163

Culpeper, J. (1996). A comparative study done by male and female hate speech comments on Nikita Mirzani's Instagram. https://repository.uhn.ac.id/bitstream/handle/123456789/9658/RESTU%20BERKA%20SIAHAAN.pdf?sequence=1&isAllowed=y

Darley, J. M., & Latane, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology*, *8*(4, Pt.1), 377–383. https://doi.org/10.1037/h0025589

Davidovic, A., Talbot, C., Hamilton-Giachritsis, C., & Joinson, A. (2023). To intervene or not to intervene: young adults' views on when and how to intervene in online harassment. *Journal of Computer-Mediated Communication*, *28*(5). https://doi.org/10.1093/jcmc/zmad027

Davidson, S. S., Hoppock, A. B., Rohmeyer, R. A., Keebler, J., & Frederick, C. M. (2020). Deindividuation in anonymous social media: Does anonymous social media lead to an increase in Non-Normative Behavior? *Scholarly Commons.* https://commons.erau.edu/publication/1414?utm_source=commons.erau.edu%2Fpublication%2F1414&utm_medium=PDF&utm_campaign=PDFCoverPages

Dillon, K. P., & Bushman, B. J. (2014). 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

Doosje, B., Ellemers, N., & Spears, R. (1995). Perceived intragroup variability as a function of group status and identification. *Journal of Experimental Social Psychology*, *31*(5), 410–436. https://doi.org/10.1006/jesp.1995.1018

Eklund, L., Von Essen, E., Jonsson, F., & Johansson, M. (2021). Beyond a dichotomous understanding of online anonymity: bridging the macro and micro level. *Sociological Research Online*, *27*(2), 486–503. https://doi.org/10.1177/13607804211019760

El-Shinnawy, M., & Vinze, A. S. (1997). Technology, culture and persuasiveness: a study of choice-shifts in group settings. *International Journal of Human-Computer Studies*, *47*(3), 473–496. https://doi.org/10.1006/ijhc.1997.0138

Everett, J. a. C., Faber, N. S., & Crockett, M. (2015). Preferences and beliefs in ingroup favoritism. *Frontiers in Behavioral Neuroscience*, *9*. https://doi.org/10.3389/fnbeh.2015.00015

Fan, C. A., Hara-Hubbard, K. K., Barrington, W. E., & Baquero, B. (2022). The experience of hate incidents across racial and ethnic groups during the COVID-19 pandemic. *Frontiers in Public Health*, *10*. https://doi.org/10.3389/fpubh.2022.982029

Finstad, F. (2010). Response interpolation and scale sensitivity. *Journal of Usability Studies Archive*. https://doi.org/10.5555/2835434.2835437

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

Fulantelli, G., Taibi, D., Scifo, L., Schwarze, V., & Eimler, S. C. (2022). Cyberbullying and Cyberhate as two interlinked instances of Cyber-Aggression in Adolescence: A Systematic review. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.909299

Gahagan, K., Vaterlaus, J. M., & Frost, L. R. (2015). College student cyberbullying on social networking sites: Conceptualization, prevalence, and perceived bystander responsibility. *Computers in Human Behavior*, *55*, 1097–1105. https://doi.org/10.1016/j.chb.2015.11.019

Gillespie, T. (2019). Custodians of the internet. In *Yale University Press eBooks*. https://doi.org/10.12987/9780300235029

Halim, S. M., Irtiza, S., Hu, Y., Khan, L., & Thuraisingham, B. (2023). WokeGPT: Improving Counterspeech Generation Against Online Hate Speech by Intelligently Augmenting Datasets Using a Novel Metric. *2023 International Joint Conference on Neural Networks, 1-10 IEEE*. https://doi.org/10.1109/ijcnn54540.2023.10191114

Harmon, R. R., & Coney, K. A. (1982). The persuasive effects of source credibility in buy and lease situations. *Journal of Marketing Research*, *19*(2), 255–260. https://doi.org/10.1177/002224378201900209

Hayes, A. F. (2022). *Introduction to mediation, moderation, and conditional process analysis: A Regression-Based Approach*. Guilford Publications.

Hite, D. M., Voelker, T., & Robertson, A. (2014). Measuring Perceived Anonymity: the development of a context independent instrument. *Journal of Methods and*

Measurement in the Social Sciences, *6*(1). https://doi.org/10.2458/azu_jmmss.v5i1.18305

Hooi, R., & Cho, H. (2014). Avatar-driven self-disclosure: The virtual me is the actual me. *Computers in Human Behavior*, *39*, 20–28. https://doi.org/10.1016/j.chb.2014.06.019

Hovland, C. I., & Weiss, W. (1951). The influence of source credibility on communication effectiveness. *Public Opinion Quarterly*, *15*(4), 635. https://doi.org/10.1086/266350

Jahan, M. S., & Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, *546*, 126232. https://doi.org/10.1016/j.neucom.2023.126232

Jenni, K., & Loewenstein, G. (1997). Explaining the Identifiable Victim Effect. *Journal of Risk and Uncertainty*, *14*(3), 235–257. https://doi.org/10.1023/a:1007740225484

Jeyagobi, S., Munusamy, S., Kamaluddin, M. R., Badayai, A. R. A., & Kumar, J. (2022). Factors influencing negative cyber-bystander behavior: A systematic literature review. *Frontiers in Public Health*, *10*. https://doi.org/10.3389/fpubh.2022.965017

Jia, Y., & Schumann, S. (2023). Tackling hate speech online: The effect of counter-speech on subsequent bystander reactions. *London University*. https://doi.org/10.33767/osf.io/9jmza

Karasavva, V., & Mikami, A. (2024). I'll be there for you? The bystander intervention model and cyber aggression. *Cyberpsychology Journal of Psychosocial Research on Cyberspace*, *18*(2). https://doi.org/10.5817/cp2024-2-1

Koehler, C., & Weber, M. (2018). "Do I really need to help?!" Perceived severity of cyberbullying, victim blaming, and bystanders' willingness to help the victim. *Cyberpsychology Journal of Psychosocial Research on Cyberspace*, *12*(4). https://doi.org/10.5817/cp2018-4-4

Lakens, D. (2022). Sample size justification. *Collabra Psychology*, *8*(1). https://doi.org/10.1525/collabra.33267

Latané, B., & Darley, J. M. (1970). The unresponsive bystander: Why Doesn't He Help*?* *Appleton-Century-Crofts*.

Leonhard, L., Rueß, C., Obermaier, M., & Reinemann, C. (2018). Perceiving threat and feeling responsible. How severity of hate speech, number of bystanders, and prior reactions of others affect bystanders' intention to counterargue against hate speech on Facebook. *Studies in Communication and Media*, *7*(4), 555–579. https://doi.org/10.5771/2192-4007-2018-4-555

Lesner, T. H., & Rasmussen, O. D. (2014). The identifiable victim effect in charitable giving: evidence from a natural field experiment. *Applied Economics*, *46*(36), 4409–4430. https://doi.org/10.1080/00036846.2014.962226

Levine, M., Cassidy, C., Brazier, G., & Reicher, S. (2002). Self-Categorization and Bystander Non-intervention: two experimental studies1. *Journal of Applied Social Psychology*, *32*(7), 1452–1463. https://doi.org/10.1111/j.1559-1816.2002.tb01446.x

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

Liebst, L. S., Philpot, R., Bernasco, W., Dausel, K. L., Ejbye-Ernst, P., Nicolaisen, M. H., & Lindegaard, M. R. (2019). Social relations and presence of others predict bystander intervention: Evidence from violent incidents captured on CCTV. *Aggressive Behavior*, *45*(6), 598–609. https://doi.org/10.1002/ab.21853

Lucassen, T., & Schraagen, J. M. (2010). Trust in Wikipedia. *WICOW '10: Proceedings of the 4th Workshop on Information Credibility*, 19–26. https://doi.org/10.1145/1772938.1772944

Marx, G. T. (1999). What's in a name? Some reflections on the sociology of anonymity. *The Information Society*, *15*(2), 99–112. https://doi.org/10.1080/019722499128565

Mathew, B., Dutt, R., Goyal, P., & Mukherjee, A. (2019). Spread of Hate Speech in Online Social Media. *Proceedings of the 10th ACM Conference on Web Science*, 173–182. https://doi.org/10.1145/3292522.3326034

Mathew, B., Saha, P., Tharad, H., Rajgaria, S., Singhania, P., Maity, S. K., Goyal, P., & Mukherje, A. (2018). Thou shalt not hate: Countering Online Hate Speech. *Cornell University*. https://doi.org/10.48550/arxiv.1808.04409

McLaughlin, C., & Vitak, J. (2011). Norm evolution and violation on Facebook. *New Media & Society*, *14*(2), 299–315. https://doi.org/10.1177/1461444811412712

Metzger, M. J., Flanagin, A. J., & Medders, R. B. (2010). Social and heuristic approaches to credibility evaluation online. *Journal of Communication*, *60*(3), 413–439. https://doi.org/10.1111/j.1460-2466.2010.01488.x

Nickerson, C. (2023, October 24). *Deindividuation in Psychology: Definition & Examples*. Simply Psychology. https://www.simplypsychology.org/what-is-deindividuation.html

Obermaier, M. (2022). Youth on standby? Explaining adolescent and young adult bystanders' intervention against online hate speech. *New Media & Society*, *26*(8), 4785–4807. https://doi.org/10.1177/14614448221125417

Osama, Y. (2023). Countering Online Hate Speech: A Qualitative study of digital Literacy training programs. *EJSC*, *2023*(84), 189–240. https://ejsc.journals.ekb.eg/article_325444_e9e2685848ffe8eb2953f9aee976b61a.pdf

Pabian, S., & Vandebosch, H. (2023). The Dark Tetrad, online moral disengagement, and online aggression perpetration among adults. *Telematics and Informatics Reports*, *11*, 100089. https://doi.org/10.1016/j.teler.2023.100089

Pabian, S., Vandebosch, H., Poels, K., Van Cleemput, K., & Bastiaensens, S. (2016). Exposure to cyberbullying as a bystander: An investigation of desensitization effects among early adolescents. *Computers in Human Behavior*, *62*, 480–487. https://doi.org/10.1016/j.chb.2016.04.022

Pluta, A., Mazurek, J., Wojciechowski, J., Wolak, T., Soral, W., & Bilewicz, M. (2023). Exposure to hate speech deteriorates neurocognitive mechanisms of the ability to understand others' pain. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-31146-1

Quirk, R., & Campbell, M. A. (2014). On standby? A comparison of online and offline witnesses to bullying and their bystander behaviour. *Educational Psychology*, *35*(4), 430–448. https://doi.org/10.1080/01443410.2014.893556

Qureshi, K. A., & Sabih, M. (2021). Un-Compromised Credibility: Social Media Based Multi-Class Hate Speech Classification for text. *IEEE Access*, *9*, 109465–109477. https://doi.org/10.1109/access.2021.3101977

Rains, S. A. (2007). The anonymity effect: The influence of anonymity on perceptions of sources and information on health websites. *Journal of Applied Communication Research*, *35*(2), 197–214. https://doi.org/10.1080/00909880701262666

Rains, S. A., & Scott, C. R. (2007). To identify or not to identify: A theoretical model of receiver responses to anonymous communication. *Communication Theory*, *17*(1), 61–91. https://doi.org/10.1111/j.1468-2885.2007.00288.x

Rashidi, Y., Kapadia, A., Nippert-Eng, C., & Su, N. M. (2020). "It's easier than causing confrontation": Sanctioning Strategies to Maintain Social Norms and Privacy on

Social Media. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW1), 1–25. https://doi.org/10.1145/3392827

Rennstam, J. (2017). Control. *The International Encyclopedia of Organizational Communication*, 1–22. https://doi.org/10.1002/9781118955567.wbieoc044

Reynolds, H., Tseung-Wong, C. N., & Kelty, S. F. (2023). Bystander intervention in coercive control: Do ethnic identity and acceptance of coercive control influence willingness to intervene? *Journal of Interpersonal Violence*, *39*(5–6), 1082–1103. https://doi.org/10.1177/08862605231212177

Rim, S., Amit, E., Fujita, K., Trope, Y., Halbeisen, G., & Algom, D. (2014). How words transcend and pictures immerse. *Social Psychological and Personality Science*, *6*(2), 123–130. https://doi.org/10.1177/1948550614548728

Rutter, M. (1987). Psychosocial resilience and protective mechanisms. *American Journal of Orthopsychiatry*, *57*(3), 316–331. https://doi.org/10.1111/j.1939-0025.1987.tb03541.x

Salminen, J., Hopf, M., Chowdhury, S. A., Jung, S., Almerekhi, H., & Jansen, B. J. (2020). Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences*, *10*(1). https://doi.org/10.1186/s13673-019-0205-6

Salmivalli, C. (2009). Bullying and the peer group: A review. *Aggression and Violent Behavior*, *15*(2), 112–120. https://doi.org/10.1016/j.avb.2009.08.007

Salmivalli, C., Lagerspetz, K., Bjorkqvist, K., Osterman, K., & Kaukiainen, A. (1996). Bullying as a group process: participant roles and their relations to social status within the group. *Aggressive Behavior*, *22*(1), 1–15. https://onlinelibrary.wiley.com/doi/epdf/10.1002/(SICI)1098-2337(1996)22%3A1%3C1%3A%3AAID-AB1%3E3.0.CO%3B2-T

Schäfer, S., Sülflow, M., & Reiners, L. (2022). Hate speech as an indicator for the state of the society. *Journal of Media Psychology Theories Methods and Applications*, *34*(1), 3–15. https://doi.org/10.1027/1864-1105/a000294

Schmid, U. K., Kümpel, A. S., & Rieger, D. (2022). How social media users perceive different forms of online hate speech: A qualitative multi-method study. *New Media & Society*, *26*(5), 2614–2632. https://doi.org/10.1177/14614448221091185

Schöne, J. P., Garcia, D., Parkinson, B., & Goldenberg, A. (2023). Negative expressions are shared more on Twitter for public figures than for ordinary users. *PNAS Nexus*, *2*(7). https://doi.org/10.1093/pnasnexus/pgad219

Siahaan, R. B. (2023). Comparative study done by male and female hate speech comments on Nikita Mirzani's Instagram. https://repository.uhn.ac.id/handle/123456789/9658

Soral, W., Bilewicz, M., & Winiewski, M. (2017). Exposure to hate speech increases prejudice through desensitization. *Aggressive Behavior*, *44*(2), 136–146. https://doi.org/10.1002/ab.21737

Suler, J. (2004). The online disinhibition effect. *CyberPsychology & Behavior*, *7*(3), 321–326. https://doi.org/10.1089/1094931041291295

Sweeney, M. S. (2003). Censorship. In *Elsevier eBooks* (pp. 189–204). https://doi.org/10.1016/b0-12-387670-2/00024-8

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

Vilanova, F., Beria, F. M., Costa, Â. B., & Koller, S. H. (2017). Deindividuation: From Le Bon to the social identity model of deindividuation effects. *Cogent Psychology*, *4*(1), 1308104. https://doi.org/10.1080/23311908.2017.1308104

Wagenknecht, T., Teubner, T., & Weinhardt, C. (2016). The impact of anonymity on communication persuasiveness in online participation. *International Conference on Information Systems*, 1. https://aisel.aisnet.org/icis2016/HumanBehavior/Presentations/1/

Walker, J. A., & Jeske, D. (2016). Understanding bystanders' willingness to intervene in traditional and cyberbullying scenarios. *International Journal of Cyber Behavior Psychology and Learning*, *6*(2), 22–38. https://doi.org/10.4018/ijcbpl.2016040102

Walther, J. B. (2015). Social Information Processing Theory (CMC). *The International Encyclopedia of Interpersonal Communication*, 1–13. https://doi.org/10.1002/9781118540190.wbeic192

Wang, S. (2020). The influence of anonymity and incivility on perceptions of user comments on news websites. *Mass Communication & Society*, *23*(6), 912–936. https://doi.org/10.1080/15205436.2020.1784950

Warner, M., & Wang, V. (2019). Self-censorship in social networking sites (SNSs) – privacy concerns, privacy awareness, perceived vulnerability and information management. *Journal of Information Communication and Ethics in Society*, *17*(4), 375–394. https://doi.org/10.1108/jices-07-2018-0060

Weidlich, J., Kreijns, K., Rajagopal, K., & Bastiaens, T. (2018, June 25). *What Social Presence is, what it isn't, and how to measure it: A work in progress*. Learning & Technology Library (LearnTechLib). https://www.learntechlib.org/primary/p/184456/

Wilson, R. A., & Land, M. (2021). *Hate speech on social media: content moderation in context*. Digital Commons @ UConn. Retrieved September 9, 2024, from https://opencommons.uconn.edu/law_papers/535?utm_source=opencommons.uconn.edu%2Flaw_papers%2F535&utm_medium=PDF&utm_campaign=PDFCoverPages

Wong, R. Y. M., Cheung, C. M. K., & Xiao, B. (2016). Combating online abuse: What drives people to use online reporting functions on social networking sites. *Hawaii International Conference on System Sciences (HICSS)*, *49*, 415–424. https://doi.org/10.1109/hicss.2016.58

Woods, F. A., & Ruscher, J. B. (2021). Viral sticks, virtual stones: addressing anonymous hate speech online. *Patterns of Prejudice*, *55*(3), 265–289. https://doi.org/10.1080/0031322x.2021.1968586

Young, R., Miles, S., & Alhabash, S. (2018). Attacks by Anons: a content analysis of aggressive posts, victim responses, and bystander interventions on a social media site. *Social Media + Society*, *4*(1), 205630511876244. https://doi.org/10.1177/2056305118762444

Yu, C. (2024, July 10). *Online Hate Speech Detection and Management from Bystander Intervention Perspective based on ETPB Model*. https://journal.esrgroups.org/jes/article/view/5156

Zapata, J., Sulik, J., Von Wulffen, C., & Deroy, O. (2024). Bystanders' collective responses set the norm against hate speech. *Humanities and Social Sciences Communications*, *11*(1). https://doi.org/10.1057/s41599-024-02761-8

Zhang, Z., & Luo, L. (2019). Hate speech detection: A solved problem? The challenging case of long tail on Twitter. *Semantic Web*, *10*(5), 925–945. https://doi.org/10.3233/sw-180338

Zhao, H., Xu, Y., Li, L., Liu, J., & Cui, F. (2024). The neural mechanisms of identifiable victim effect in prosocial decision-making. *Human Brain Mapping*, *45*(2). https://doi.org/10.1002/hbm.26609

Appendix A

Pretest

To ensure the most suitable stimulus material, a pretest was conducted in which 5 different comments were shown to 7 respondents who were not participating in the study. Participants in the pretest were asked to assess the perceived threat and the perceived realism (Cho et al., 2012) of these 5 comments, on a 7-point Likert scale.

1. "Body positivity is gewoon een excuus voor walgelijke varkens om zichzelf niet te veranderen. Je bent niet 'mooi', je bent gewoon vet en lui!"

"Body positivity is just an excuse for disgusting pigs not to change themselves. You're not 'beautiful,' you're just fat and lazy!"

2. "Hoe kun je trots zijn op het feit dat je je lichaam kapot maakt?"

"How can you be proud of destroying your body?"

3. "Body positivity? Wat een trieste manier om jezelf beter te laten voelen over je morbide overgewicht."

"Body positivity? What a sad way to make yourself feel better about your morbid obesity."

4. "Al die zogenaamde body positivity-mensen zouden uit de sociale media moeten worden gegooid. Ze promoten letterlijk een ziekte! Fat acceptance? Wat een grap!"

"All these so-called body positivity people should be thrown out of social media. They are literally promoting a disease! Fat acceptance? What a joke!"

5. "Mensen moeten ophouden 'eigenliefde' te gebruiken als excuus om gewoon dik en lui te zijn."

"People should stop using 'self-love' as an excuse to just be fat and lazy."

Table 5

Pre-test standard descriptives on a 7-point scale

	Comment 1	Comment 2	Comment 3	Comment 4	Comment 5
Perceived threat	5.27 (*SD* = 1.07)	4.52 (*SD* = 1.71)	4.78 (*SD* = 1.48)	5.27 (*SD* = 1.38)	4.68 (*SD* = 1.40)
Perceived realism	5.80 (*SD* = 1.29)	6.26 (*SD* = 0.76)	6.09 (*SD* = 0.75)	5.31 (*SD* = 1.35)	6.09 (*SD* = 0.59)

Table 6

Pre-test perceived anonymity of aggressors on a 7-point scale

Perceived anonymity	Identifiable	Pseudonymous	Anonymous
Mean	3.00	4.80	6.09
SD	2.27	1.50	1.18

Appendix B

Tweet used in experiment

Figure 6

Stimulus material experiment

Note. The pictures were collected from the open database of Canva.

English translation:

Growing numbers of people seem to be using “body positivity” to dismiss health risks for morbid obesity. Experts say this blurs the line between self-acceptance and glorifies an unhealthy lifestyle. More and more influencers are glorifying obesity.

Growing numbers of influencers glorify obesity. Self-acceptance or new danger?

Appendix C

Scales used in experiment

The experiment started with the measurement of *indirect* intention to intervene. Items used for the scale were adopted from Jia & Schumann (2023) and Obermaier (2022) to ensure all available intervention methods for bystanders, using a 5-point Likert scale.

<u>If you would see this comment section on this news article on your own phone, how likely is that you would respond as described below?</u>

1. I would report the profile of hate speech aggressor to the X platform

   Ik zou het profiel van de hate speech-verzender rapporteren bij het X-platform.

2. I would report the hate speech comment tot the X platform

   Ik zou de hate speech comment rapporteren bij het X-platform.

3. I would like other comment(s) that contradict the hate speech in the comment section

   Ik zou andere comment(s) liken die de hate speech in de commentsectie tegenspreken (up-voting).

4. Reacting with non-verbal cues like dislike of sad emoji to signal disapproval.

   Ik zou commenten met non-verbale signalen zoals een 'dislike' of een verdrietige emoji's om negatieve mening aan te geven.

Secondly, the experiment followed with the measurement of *direct* intention to intervene. Items used for the scale were adopted from Jia & Schumann (2023) and Obermaier (2022) to ensure all available intervention methods for bystanders, using a 5-point Likert scale.

<u>If you would see this comment section on this news article on your own phone, how likely is that you would respond as described below?</u>

1. Condemning the hate speech with factual contradicting argumentation

   Ik zou de hate speech tegenspreken met feitelijke, tegenstrijdige argumenten.

2. Responding to hate speech with a question

   Ik zou reageren op hate speech met een vraag.

3. Sanctioning the communication by reminding to politeness

   Ik zou de deze manier van communiceren veroordelen door de hate speech verzender te herinneren aan beleefdheid.

4. Educating the hate speech aggressor

   Ik zou proberen om de hate speech-aanvaller meer informatie te geven over de mensen benoemd in het nieuwsartikel.

5. Trying to persuade the hate speech aggressor to change opinion

   Ik zou proberen de hate speech-aanvaller te overtuigen om van mening te veranderen.

6. Comment to show support and empathy for victims of hate speech

   Ik zou een reactie plaatsen om steun en empathie voor de slachtoffers van de hate speech te tonen.

7. Comment to gather more support for victim of hate speech

   Ik zou een reactie plaatsen om meer steun voor de mensen genoemd in het nieuwsartikel te verzamelen.

8. Responding to content of hate speech with similar offensive words

   Ik zou reageren op de inhoud van de hate speech met vergelijkbare beledigende woorden.

9. Responding to the aggressor of hate speech with hate speech

   Ik zou reageren op de hate speech-verzender met hate speech.

10. Insulting the hate speech aggressor

    Ik zou de hate speech verzender beledigen.

11. Confront the speech aggressor in a direct message

    Ik zou de hate speech aanvaller confronteren in een private message.

12. Support the victim in a direct message

    Ik zou het slachtoffer steunen in een private message.

Thirdly, the measurement of the perceived threat was conducted. This scale was constructed for this experiment, using a 7-point Likert scale.

Please indicate your level of agreement or disagreement with each of the following statements.

1. Confronting online aggressors could lead to offline threats or harassment.

   Het confronteren van de hate speech verzender kan leiden tot offline bedreigingen en pesterijen.

2. I feel personally distressed when I see someone being attacked in this way

   Ik krijg een naar gevoel wanneer ik comments lees waarin mensen op deze manier worden aangevallen.

3. I am unsure about the aggressor's intentions when their identity is hidden

   Ik weet niet wat de intenties van de agressor zijn wanneer hun identiteit verborgen is.

4. I feel that intervening directly could make me a target for retaliation.

   Ik heb het gevoel dat direct ingrijpen mij een doelwit kan maken voor hate speech.

5. It is difficult to predict how the person posting hate speech might react in the future.

   Het is moeilijk te voorspellen hoe de persoon die hate speech plaatst in de toekomst zal reageren.

6. Speaking up against online hate speech could negatively impact my reputation.

   Als ik deze hate speech tegenspreek, kan dit mijn reputatie negatief beïnvloeden.

7. Hateful comments create an online environment that feels threatening.

   Haatdragende opmerkingen creëren een online omgeving die bedreigend aanvoelt

8. The actions of those posting hate speech make me feel uneasy about intervening

   De acties van degenen die hate speech plaatsen, geven mij een te ongemakkelijk gevoel om in te grijpen

Fourthly, social presence was measured by including the following statements from the social presence scale, using a 7-point Likert scale. (Weidlich et al., 2018).

Please indicate your level of agreement or disagreement with each of the following statements.

1. In this online environment, it feels as if I read a comment from a 'real' person and not with an abstract anonymous person.

   De hate speech verzender voelde voor mij als een 'echt' persoon.

2. I could not form clear impressions of the victim.* - recoded.

   Ik kon geen duidelijke indruk vormen van de hate speech verzender. * - recoded.

3. The victim felt so 'real' that I almost believed that we were not virtual at all.

   De hate speech verzender voelde 'echt' voor mij.

4. I imagine that I really can see hate speech aggressor to be in front of me.
   Ik kon me voorstellen dat ik de hate speech verzender voor me zag staan.

5. It feels as if the hate speech aggressor and me are in the same room.
   Het voelt alsof de hate speech verzender en ik in dezelfde ruimte zijn

6. strongly feel the presence of the hate speech aggressor.
   Ik voelde sterk de aanwezigheid van de hate speech verzender.

Fifthly, perceived credibility was measured by including the following statements from the perceived credibility scale, using a 7-point Likert scale (Chesney & Su, 2010).

Please indicate your level of agreement or disagreement with each of the following statements.

1. I perceive the hate speech aggressor to be credible
   Ik ervaar de hate speech-aanvaller als geloofwaardig.

2. I perceive the hate speech aggressor to be accountable for their message
   Ik ervaar de hate speech-aanvaller als verantwoordelijk voor zijn/haar bericht.

3. I perceive the hate speech aggressor to be trustworthy
   Ik ervaar de hate speech-aanvaller als betrouwbaar.

4. I perceive the information to be believable
   Ik ervaar de informatie als geloofwaardig.

5. I perceive the information to be accurate
   Ik ervaar de informatie als nauwkeurig.

6. I perceive the information to be trustworthy
   Ik ervaar de informatie als betrouwbaar.

7. I perceive the information to be complete
   Ik ervaar de informatie als volledig.

After measuring the mediating variables, the first control variable was measured. Perceived anonymity was measured by including items from the perceived anonymity scale (Hite et al., 2014) and adjusted to perceived anonymity context from the perspective of the bystander instead of the aggressor. This scale was measured on a 7-point Likert scale.

<u>Please indicate your level of agreement or disagreement with each of the following statements.</u>

1. I am confident that I could identify the hate speech aggressor personally
   Ik ben ervan overtuigd dat ik dit profiel persoonlijk zou kunnen identificeren

2. I believe that the personal identity of the hate speech aggressor is unknown to me *rec
   *Ik geloof dat ik de persoonlijke identiteit van dit profiel NIET herleidbaar voor mij is**

3. The hate speech aggressor is easily identified as an individual
   Dit profiel is makkelijk om persoonlijk te identificeren.

4. Other people like me are likely to be able to identify the hate speech aggressor
   Andere mensen zoals ik zijn waarschijnlijk in staat om dit profiel te identificeren

5. The personal identity of the hate speech aggressor is known to others
   De persoonlijke identiteit van dit profiel is bekend bij anderen.

The second control variable, perceived similarity, was measured by including items from the group identification scale (Doosje et al., 1995). The items were adjusted to fit the context of the body-positivity movement, using a 7-point Likert scale.

Please indicate your level of agreement or disagreement with each of the following statements.

1. I identify with the body-positivity movement.
   Ik identificeer mij met de body-positivity beweging.

2. I see myself as a member of the body-positivity movement.
   Ik lijk op een lid van de body-positivity beweging.

3. I do not feel a connection with the body-positivity movement. *****
   *Ik voel GEEN connectie met de body-positivity beweging. **

4. I feel strong ties with the body-positivity movement.
   Ik voel mij verbonden met de body-positivity beweging.

5. I belong to the same social group as the body-positivity movement.
   Ik behoor tot dezelfde sociale groep als de body-positivity beweging.