MUNI
FSS

JOURNAL OF PSYCHOSOCIAL RESEARCH ON CYBERSPACE
CYBERPSYCHOLOGY

Pabian, S. (2025). Bystanders' perceptions on online hate speech: Investigating the effects of perpetrators' justifications and the bystander's role on bystanders' attitude and prosocial intervention intentions. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace, 19*(4), Article 6.
https://doi.org/10.5817/CP2025-4-6

Bystanders' Perceptions on Online Hate Speech: Investigating the Effects of Perpetrators' Justifications and the Bystander's Role on Bystanders' Attitude and Prosocial Intervention Intentions

Sara Pabian

Tilburg center for Cognition and Communication, Tilburg University, Tilburg, Netherlands
Department of Communication Studies, University of Antwerp, Antwerp, Belgium

Abstract

On social media, users are exposed to online hate speech (OHS), which is a type of speech that attacks a person or a group based on a group characteristic, e.g., gender identity or sexual orientation. Not every bystander evaluates OHS as offensive and/or feels the need to intervene, which can lead to the continuation of OHS and damaging consequences for victims. The goal of the present study was to understand attitudinal and behavioral components of bystanders' perceptions on OHS by investigating content-related, contextual, and personal characteristics. More precisely, the effects of the presence or absence of online moral disengagement strategies or moral excuses in OHS messages (e.g., "I'm posting this because it doesn't hurt if I share my opinion online") and the bystander's role (pure bystander or vicarious victim) on bystanders' attitudes and behaviors were tested, while controlling for previous experience with OHS and connectedness with the target group. To this aim, a repeated measures experiment (5x2x2 mixed design) was conducted among 633 adults aged 18–25. The results indicated no difference in bystanders' perceived offensiveness of OHS and intention to intervene when exposed to OHS containing a moral excuse compared to OHS without. When bystanders were vicarious victims (being exposed to OHS targeting an individual with whom the bystander shares the targeted group characteristic), OHS was perceived as more offensive and bystanders had a higher intention to intervene with prosocial bystander behavior, compared to when bystanders did not share the group characteristic. Theoretical and practical implications are discussed.

Keywords: social media; online hate speech; bystanders; attitudes; interventions

Editorial Record

First submission received:
September 8, 2024

Revisions received:
March 14, 2025
July 2, 2025

Accepted for publication:
July 16, 2025

Editor in charge:
Lenka Dedkova

Introduction

Much to the concern of scholars, field experts, and social media users, the amount of detected online hate speech has been rising (e.g., Paschalides et al., 2020). Moreover, researchers have found an increase in self-reported exposure rates to online hate speech in surveys among social media users, especially among young adults (e.g., Obermaier, 2024; Reichelmann et al., 2021). Online hate speech (OHS) is defined as the expression towards an individual or group of "hatred or degrading attitudes toward a collective" (Hawdon et al., 2017, p. 254), aiming to devalue and demean them collectively, whereby the collective has a shared characteristic, such as race, religion, ethnic origin, gender identity, disability, or sexual orientation (Johnson et al., 2019). Hate speech can target every possible group that has a group characteristic in common (Paasch-Colberg et al., 2021). Recent studies, including systematic reviews, have highlighted the need for examinations on the side of the audience: those who are not directly targeted but who are exposed to OHS while using social media (e.g., Bormann et al., 2022; Matamoros-Fernández & Farkas, 2021). Researching this large group, also called bystanders or witnesses of OHS, is important as their reactions have a pivotal role in the continuation of OHS and the impact of OHS for victims, perpetrators, and society as a whole (Kümpel & Rieger, 2019). However, it remains unclear how bystanders perceive OHS (Schmid et al., 2024).

There are indications that there are differences in perceptions: some recognize OHS and decide to look at the post, others not; some perceive OHS as severe, disturbing, offensive, and/or harmful (attitudinal component), others not; and some perceive the need to intervene (behavioral component), for instance by reporting OHS to service providers, while others not (Costello et al., 2019; Kenski et al., 2020; Ohme & Mothes, 2020; Schmid et al., 2024). Following media perception theories, perceptions of OHS might hinge on the presentation of the message, contextual factors, and the characteristics of the receiver (Kenski et al., 2020; Schmid et al., 2024). In this article, we argue that bystanders' attitudinal and behavioral perceptions of OHS in user comments can depend on the presence or absence of moral disengagement strategies in OHS messages (content-related characteristic) and the bystander's role (pure bystander or vicarious victim; contextual determinant), while controlling for the bystander's previous experience with OHS (personal characteristic) and bystander's personal connectedness with the target group (personal characteristic). An understanding of these perceptions is necessary to move the field forward to developing effective interventions aimed at, for instance, raising awareness or promoting appropriate bystander behavior.

Bystanders' Perceptions When Witnessing OHS

A number of studies have investigated attitudinal and/or behavioral components of bystanders' perceptions. Those that investigated attitudinal components have, for instance, looked at bystanders' perceived offensiveness of OHS (Kümpel & Unkel, 2023), incivility (Obermaier et al., 2023), harmfulness to society (Kümpel & Unkel, 2023), and hatefulness (Papcunová et al., 2023). Those studies that focused on behavioral components have investigated different reactions of bystanders, including general intention to intervene (Costello et al., 2023; Obermaier, 2024), intention to engage in constructive counterspeech (Kunst et al., 2021; Leonhard et al., 2018; Obermaier, 2024), reporting behavior (Mohseni, 2023; Naderer, et al., 2023; Obermaier, 2024; Wilhelm et al., 2020), and flagging OHS (Kunst et al., 2021; Obermaier, 2024). However, research indicates that a large proportion of the bystanders of OHS do nothing when they are exposed to OHS (e.g., 42%; Obermaier, 2024). Bystanders that do decide to react can behave in different ways. For instance, they can report the OHS message to the platform on which it was taking place, or they can engage in counterspeech by, for instance, sharing a disagreeing comment in public or via a private message to the victim and/or perpetrator. The present study will focus on prosocial behaviors, such as the behaviors mentioned above and will not explore antisocial bystander behaviors. Examples of the latter are liking the OHS message (joining in) or destructive counterspeech, such as offending the perpetrator of OHS (Obermaier, 2024).

A large amount of literature focusing on understanding whether bystanders intervene or not intervene when witnessing aggression or violence have built on Latané and Darley's (1970) bystander intervention model. The latter suggests that if the number of bystanders observing an aggressive act increases, the likelihood that individuals will intervene decreases. Latané and Darley's five-step model proposes that, first, bystanders must notice the act; second, they have to assess the situation as serious; third, they must feel responsible to do something; fourth, they have to (be able) to reflect on how to help; and, finally, they need to decide to intervene and actually intervene. If all steps are completed, a bystander will intervene or do something. The five-step model has been discussed and empirically tested, with mixed findings, in online contexts, including among bystanders of OHS (e.g., Leonhard et al., 2018; Obermaier et al., 2023).

A number of authors have indicated the need to integrate various content-related, contextual, and personal factors besides the components highlighted in the bystander intervention model to understand whether bystanders intervene or not and, if they intervene, which behavior(s) they perform (Obermaier, 2024; Rudnicki et al., 2022). Authors have also indicated that the bystander intervention model might be less suitable for predicting bystander behaviors that are not performed in public, such as reporting OHS to service providers or sending private messages, as bystanders do not have to fear negative evaluation of other bystanders in reaction to their bystander behavior (Leonhard et al., 2018). Also, for attitudinal components of bystanders' perceptions, researchers have indicated the need to investigate (a diverse set of) content-related, contextual, and personal determinants (Costello et al., 2023; Papcunová et al., 2023).

Content-Related Characteristics of Attitudinal and Behavioral Components of OHS Bystanders' Perceptions

The present study wants to contribute to the recent focus in OHS studies on content characteristics (e.g., Kümpel & Unkel, 2023; Obermaier et al., 2023; Schmid, 2023; Wilhelm et al., 2020) to understand attitudinal and behavioral components of bystanders' perceptions. For instance, research has shown that OHS containing more severe elements, such as threats of violence, increases bystanders' willingness to intervene with counterspeech if OHS is considered as threatening and bystanders feel responsible to act (Leonhard et al., 2018). Another study showed that humorously portrayed hate speech in memes is less often viewed as hostile than non-humorous hate speech (Schmid, 2023). This was especially true for more implicit forms of OHS without overt declarations of hatred, such as negative stereotyping. Research has suggested that not only the type of OHS (e.g., implicit/explicit, humorous OHS) can influence attitudinal and behavioral components of bystanders' perceptions, also the presence of justifications or excuses for performing OHS might influence bystanders' perceptions.

The use of justifications in OHS messages has been studied by Wilhelm et al. (2020). In their study, hate comments including justifications, such as denial of negative intent (*It was just for fun*), were less likely to be reported by bystanders compared to OHS without neutralizations. In their study, the authors applied the neutralization theory of Sykes and Matza (1957). According to the theory, by applying neutralization or verbalizing justifications or excuses, perpetrators prevent themselves from feeling guilt. Moreover, due to the presence of neutralization expressed by the perpetrator, bystanders' perceived deviance of OHS might be reduced (Sykes & Matza, 1957; Wilhelm et al., 2020). In this way, perpetrators 'mask' their OHS messages, by making OHS appear moral and acceptable, and might also avoid 'punishment' or resistance, such as having their posts or account reported by other users, or receiving counterspeech. In the study of Wilhelm et al. (2020), bystanders' willingness to report OHS comments including justifications or excuses was half as low as bystanders' willingness to report OHS comments without these techniques. Researchers argue that such justifications might lead to acceptance and normalization of deviant behaviors (Henry, 2009; Wilhelm et al., 2020).

Content analyses of OHS messages indeed indicate that (some) perpetrators of OHS express justifications in their OHS messages (e.g., D'Errico & Paciello, 2018; Faulkner & Bliuc, 2016; Nurhadiyanto & Octaviani, 2021; Paciello et al., 2019). For instance, in the study of Nurhadiyanto and Octaviani (2021), three types of justifications were found in OHS messages directed at celebrities on Instagram, namely denying injury for those that are targeted, denying victimization (e.g., by indicating that those who are targeted deserve OHS), and appeals to higher loyalties. Faulkner and Bliuc (2016) found in their content analysis of racist reactions on news website articles that justifications were commonly used. More precisely, approximately 90% of the racist comments that were analyzed contained at least one justification, such as reframing the harmful action to appear moral, minimizing the perpetrator's role in causing harm, and reframing the victim as deserving the harmful action. Finally, content analyses of OHS responses on (1) a tweet from an Italian female journalist who defends the rights of immigrants and females (Paciello et al., 2019) and (2) a Facebook post in support of immigrants after a serious shipwreck causing the death of 700 people (D'Errico & Paciello, 2018) showed the presence of different types of justifications for posting OHS reactions, those that are described by Bandura (1986, 1999, see next paragraph).

The present study wants to continue and expand the experimental work that has been done on the effects of moral excuses (Wilhelm et al., 2020). The earlier mentioned neutralization theory has quite some overlap with Albert Bandura's moral disengagement strategies (Ribeaud & Eisner, 2010). According to Bandura, moral disengagement strategies are cognitive strategies that have a psychological nature and that are used by individuals to excuse their performance of immoral or harmful behavior (Bandura 1986; Marín-López et al., 2020). Whereas Bandura (1986, 1999) seems to highlight the psychological nature of these strategies, Sykes and Matza (1957) have emphasized the social nature of learning neutralization techniques (Cardwell & Copes, 2021).

The present study will focus on the perpetrators' expression of the four psychological strategies described by Bandura et al. (1999) and how these influence bystanders' perceptions: (1) moral justification or the reconstruction of the conduct itself so it is not viewed as immoral, (2) diffusion of responsibility or minimizing one's role in causing harm, (3) distortion of consequences or minimizing the consequences that result from the harmful behavior, and (4) attribution of blame or devaluing targets of harmful behaviors as human beings and blaming them for what is being done to them. These strategies have already been applied and 'translated' to the online environment by previous research, taking into account the online context and the affordances of ICT, such as the possibility to act anonymously and the paucity of social-emotional cues (Maftei et al., 2022; Runions & Bak, 2015).

In numerous survey studies on (other forms of) online aggressive behavior, perpetrators have indicated their agreement more strongly with (online) moral disengagement beliefs that excuse immoral behavior, compared to those who do not engage in online aggression (e.g., Nocera et al., 2022; Pabian & Vandebosch, 2023). The earlier cited content analyses seem to suggest that these beliefs are also present among perpetrators of OHS, as they seem to be expressed in (some of) their messages. The present study will investigate the effects of the presence of online moral disengagement strategies in OHS messages on bystanders' attitudes towards OHS and their behavioral reactions. Based on the theoretical foundations of the neutralization theory and Bandura's moral disengagement strategies and the results of Wilhelm et al. (2020), it is expected that OHS messages containing an expression of an online moral disengagement strategy will be perceived as less negative (attitudinal component of bystanders' perceptions) and will lower bystanders' intention to engage in prosocial bystander behavior (behavioral component of bystanders' perceptions). Note that, based on the literature, no expectations could be formulated about potential differences in effects of different online moral disengagement strategies to excuse OHS on bystanders' attitudes and behaviors. However, the study will compare the effects of the different strategies to explore any potential differences.

H1: OHS with an online moral disengagement strategy statement included will be (a) perceived as less negative by bystanders, and (b) associated with a lower intention to intervene with prosocial bystander behavior among bystanders, compared to OHS without an online moral disengagement strategy statement.

Contextual Determinants of Attitudinal and Behavioral Components of OHS Bystanders' Perceptions

As mentioned earlier, researchers have also highlighted the need to investigate contextual factors for understanding bystanders' perceptions and behaviors. In the present study, contextual determinants are perceived as determinants that are context-specific, that depend on the situation the bystander is in. Examples of contextual determinants that have been studied are perceptions about the behavior of other bystanders and about the evaluations of others if one would intervene (Leonhard et al., 2018; Obermaier, 2024), and the relationship between the victim and bystander (e.g., Kümpel & Unkel, 2023; Papcunová et al., 2023; Rudnicki et al., 2022). Regarding the latter, it might be important to consider here what the actual role of the bystander is. Recently, Stahel and Baier (2023) introduced 'vicarious victim' as an additional role in research on OHS, besides victims, perpetrators, and bystanders. The authors used this term to refer to bystanders of OHS that is targeting an individual and/or group with whom the bystander shares the targeted group characteristic and indicated that this role has been ignored in the literature on OHS (Stahel & Baier, 2023). Indeed, when OHS is directed to all group members with a specific characteristic, the bystander seems no longer a bystander, but can be considered as a 'vicarious victim', although OHS is not directly targeted at their profile or user account. In their study among 2,400 Swiss adults, Stahel and Baier (2023) found that 40.6% of their sample was at least once in the past 12 months a vicarious victim of OHS, whereas 56.4% was a 'pure' bystander, showing that both are common roles among internet users. However, there might be differences in perceptions on OHS when the bystander is a vicarious victim compared to a pure bystander.

Theories that explain differences in individuals' attitudes and behaviors towards others that belong or do not belong to their group might be helpful to predict perceptions of vicarious victims and pure bystanders, including the social identity theory. The social identity theory indicates that how individuals define and value themselves derives from their ties to (a specific) social category/ies (Stets & Burke, 2000; Turner et al., 1987). A social identity is a person's perception about their perceived membership of a relevant social group, but also a perception of the identity of this group and how this group differs from other social groups (Hogg & Abrams, 1988; Tajfel, 1974). Through social comparison, the process of looking for real or imagined similarities and differences, others who are perceived as similar to the self (within one's social group) are categorized as the ingroup; others who differ from the self (group) are categorized as the outgroup (Stets & Burke, 2000). A result of this process is an accentuation of the perceived similarities between members of the ingroup and an accentuation of the perceived differences between the self/in-group and outgroup members (Stets & Burke, 2000). Especially characteristics that are perceived as positive are linked to the ingroup, whereas characteristics that are perceived as negative are rather linked to outgroups. This can result in negative attitudes and behaviors toward the outgroup, including discrimination and polarization, but also helping ingroup members and refraining from helping outgroup members (Greer & Jewkes, 2005).

Also the empathy-altruism hypothesis can explain why bystanders are more likely to perceive deviant behavior towards an ingroup member as more offensive and have a higher intention to intervene, compared to an outgroup member being attacked. According to the hypothesis, when perceived self-other similarity and closeness is high, empathy increases and this will foster altruistic motivations and behaviors (Batson, 1991; Stürmer & Snyder, 2009).

Based on these theoretical frameworks, it is expected that:

H2: OHS directed towards (a member of) bystanders' ingroup will be (a) perceived as more negative by bystanders, and (b) associated with a higher intention to intervene with prosocial bystander behavior among bystanders, compared to OHS directed towards (a member of) bystanders' outgroup.

Note that previous studies investigating the relationship between the victim and bystander in the context of OHS rather focused on bystanders' personal connectedness with the group targeted by OHS, which could be perceived as an individual determinant of bystanders' perceptions.

Individual Traits as Control Variables

In the present study, individual characteristics are perceived as characteristics that are person-specific and that are not influenced by and thus are independent of the specific context a bystander is in. As indicated above, previous research focused on bystanders' personal connectedness with the group targeted by OHS. A number of studies have found an association between connectedness, operationalized as perceived similarity, closeness, and/or social distance, and attitudinal and behavioral components of bystanders' perceptions. More precisely, in the study of Papcunová et al. (2023), bystanders who felt closer to groups attacked in OHS perceived these OHS comments as more hateful. Kümpel & Unkel (2023) showed that higher perceived similarity to the group attacked in OHS consistently predicted an increase in perceived offensiveness, harm to society (both attitudinal components), and deletion intention (behavioral component). Rudnicki et al. (2022) highlighted in their systematic review of articles published between 2000 and 2020 on determinants of adult OHS and cyberbullying bystander interventions that, in order for bystanders to intervene, bystanders should feel connected to the victim.

Therefore, the present study will control for the potential influence of personal connectedness to the group attacked on attitudinal and behavioral components of bystanders' perceptions on OHS. Hereby it is assumed that one can feel (strongly) connected to members of the ingroup, but also to members of the outgroup, for instance based on positive previous interactions with these outgroup members. The same is true for feeling disconnected, a person can feel disconnected from members of the outgroup, but also from members of the ingroup, for instance based on negative previous interactions with these ingroup members.

Besides connectedness, also a number of other personal determinants of bystanders have been linked to OHS bystanders' attitudes and/or behaviors. For both attitudinal and behavioral components of bystanders' perceptions on OHS, researchers have highlighted the importance of previous exposure to OHS as a bystander and a victim. Studies indicate that individuals who see hate materials online as a bystander more frequently, find it more disturbing (Costello et al., 2019) and are more likely to intervene (Costello et al., 2023; Obermaier, 2024). For instance, Obermaier (2024) indicated that frequent exposure to OHS predicted bystanders' constructive counterarguing among young German adults, including behaviors such as refuting the OHS statements with facts and writing a comment that supported the affected group (Obermaier, 2024). However, there are also indications that some frequent social media users react less strongly on OHS and find it more normal, which can be explained by processes of desensitization and normalization (Schäfer et al., 2022; Schmid et al., 2024). In the study of Ping et al. (2025) among U.S. based adults, frequent bystanders of OHS were more reluctant to engage in counterspeech when witnessing OHS, due to, for instance, concerns for retaliation and third-party harassment.

Regarding previous exposure to OHS as a victim, research indicates that also those who have been victimized (being the target of OHS) find OHS more disturbing (Costello et al., 2019) and are more likely to intervene (Costello et al., 2023; Mohseni, 2023), for instance by engaging in counterspeech (Ping et al., 2025). Costello et al. (2023) indicate that both seeing OHS (as a bystander) and being targeted by OHS (as a victim) frequently may lead to the belief that online hate is a pervasive problem with potentially harmful consequences and this might create the urge to remedy the situation, whereas those who rarely see OHS might perceive it as a minor concern or no concern at all and do not feel the need to be involved/be concerned about the issue.

Taken together, research seems to indicate that bystanders' previous exposure to OHS might influence bystanders' attitudinal and behavioral components of their perceptions. There are some mixed findings for the relationship between the frequency of being exposed to OHS as a bystander and bystanders' attitudes and behaviors, whereas the frequency of being a victim of OHS in the past seems to be associated with more negative attitudes towards OHS and a higher intention to intervene with prosocial bystander behavior. Research in related fields, such as cyberbullying and other forms of antisocial behaviors online seem to show similar patterns (DeSmet et al., 2016; Dominguez-Hernandez et al., 2018; Henson et al., 2020; Pabian et al., 2016). Therefore, the present study will control for the potential influence of connectedness and previous involvement as a bystander and vicarious victim when investigating the potential influence of the presence or absence of moral disengagement strategies in OHS messages (content-related characteristic) and the bystander's role (pure bystander or vicarious victim; contextual determinant) on bystanders' attitude towards OHS and intention to intervene with prosocial bystander behavior.

Methods

Design and Procedure

To test the hypotheses, an online survey with a 5x2x2 mixed experimental design was conducted among Dutch-speaking young adult social media users aged 18 to 25. The present study focused on young adults as (1) researchers have indicated that young adults observe OHS more frequently than any other adult age group (e.g., Obermaier, 2024; Reichelmann et al., 2021) and (2) a large portion of the studies on bystander perceptions when witnessing OHS have focused on young adults, which enhances opportunities to compare the results with previous research (e.g., Costello et al., 2019, 2023; Hawdon et al., 2017; Henson et al., 2020; Naderer et al., 2023; Obermaier & Schmuck, 2022). Participants' perceptions were assessed after showing vignettes with hypothetical situations consisting of OHS comments on social media. Thereby, two factors were manipulated between subjects. First, the presence or absence of a statement in the OHS comment to excuse the immoral behavior (Bandura et al., 1999) was manipulated; five levels: (1) *no online moral disengagement strategy present in comment*, (2) *moral justification*, (3) *diffusion of responsibility*, (4) *distortion of consequences*, (5) *attribution of blame*. The second factor involved the bystander's role (two levels: (1) *OHS towards outgroup – bystander is a pure bystander* versus (2) *OHS towards ingroup – bystander is a vicarious victim*). Participants were randomly assigned to one of these ten conditions in which two vignettes were shown, differing on the group that was being attacked (within factor: a certain age group and a group with a certain gender identity). The vignettes were pretested (see below). The within factor was included to immediately replicate the findings.

Stimuli

Hypothetical vignettes were used to experimentally manipulate the two between-subjects factors, similar to previous experimental work on bystander behavior within the field of online aggression (e.g., Macaulay et al., 2022; Palladino et al., 2017). The vignettes with structurally similar OHS comments of comparable length were created after careful examinations of online discussions and memes on certain age groups (younger and older adults) and certain gender identities (males and females), in order to create speech close to online rhetorics that are currently present online (based on Kümpel & Unkel, 2023). OHS directed towards age groups and gender identities were chosen, as negative stereotypes are often spread about these groups, both online and offline, for instance, in memes (online) and spoken jokes (offline; age online: e.g., Fraser et al., 2022; age offline: e.g., Gullette, 2024; gender online: e.g., Castaño-Pulgarín, et al., 2021; gender offline: e.g., Lewis & Lupyan, 2020). In this way the stimuli materials presented speech that all/most of the participants could also witness in daily life, which increased the ecological validity of the experiment. All OHS comments included the same type of OHS, negative stereotyping, which is described in the literature as a common element of many hate speech definitions and qualifies as hate speech (Paasch-Colberg et al., 2021; Rieger et al., 2021). Negative stereotyping is perceived as a more indirect, subtle form in which the perpetrator 'hides' prejudices about an outgroup and/or in which the perpetrator strategically elevates the ingroup (Paasch-Colberg et al., 2021).

In order to select reliable stimulus materials, a pretest survey was conducted among 18- to 25-year-olds (*N* = 134). A description of the pretest results can be found in the Supplementary materials. Based on the pretest results, twenty vignettes were selected, these can be consulted in Appendix C. Each vignette started with a request to imagine reading on a social media platform a news message with the following headline. After showing the headline, which consisted of a news fact about a person with a certain group characteristic, the scenario described that other news readers had posted reactions on the article. One of the reactions was shown, consisting of an OHS comment.

Measures

After providing informed consent, participants were asked about their socio-demographics, social media use, and connectedness with certain groups (see below). Subsequently, participants were, after random assignment to one of the conditions, exposed to the first vignette in which a certain age group was attacked, and next to a vignette in which a group with a certain gender identity was targeted. After each vignette, participants were asked to label the OHS comment (open question) and rate it on perceived offensiveness and intention to perform prosocial bystander behavior. After answering these questions for each vignette, participants' previous experiences with OHS were assessed.

Attitudinal Components of Bystanders' Perceptions of OHS

In order to assess bystanders' attitude towards OHS two measurements were used. The following open question was used to assess how participant would *label* the OHS comment: *How would you label or describe this type of message?*. Participants answered this question for both OHS messages they were exposed to (OHS targeting a certain age group and gender identity), in order to measure the sentiment of their description, which is perceived as helpful in understanding attitudes and feelings of individuals (e.g., Catelli et al., 2023). As indicated in the literature overview, previous research that investigated attitudinal components of bystanders' perceptions towards OHS have looked at different variables, such as perceived offensiveness of OHS (Kümpel & Unkel, 2023), incivility (Obermaier et al., 2023), harmfulness to society (Kümpel & Unkel, 2023), and hatefulness (Papcunová et al., 2023), all measured with closed questions. The present study included both open and closed questions, as it was believed that the open questions could provide additional interesting insights, namely on different labels or terms young adult social media users use for referring to speech that researchers define as OHS. Under 'Data analysis', it is described how these answers were coded and used in further analyses.

Next, *perceived offensiveness* was assessed using three items: *This post is… "offensive", "hostile", "hurtful"*; each rated on a 7-point scale ranging from (1) *completely disagree* to (7) *completely agree* (Kümpel & Unkel, 2023). A mean score of the three items was calculated for each participant and vignette (Vignette A: *M* = 4.76, *SD* = 1.25, α = .83; Vignette B: *M* = 5.47, *SD* = 1.33, α = .86).

Behavioral Component of Bystanders' Perceptions of OHS

The intention to perform prosocial bystander behavior was measured with two items (based on: Macaulay, et al., 2022; Obermaier et al., 2023) that were assessed on a 7-point scale ranging from (1) c*ompletely disagree* to (7) *completely agree*. The items were: *I would report this message if I would see it on a social media platform* and *I would publicly respond to this message with a statement that contradicts this comment*. A mean score of the two items was calculated for each vignette for each participant (Vignette A: *M* = 3.30, *SD* = 1.60, α = .73; Vignette B: *M* = 3.61, *SD* = 1.70, α = .71).

Control Variables

Connectedness was assessed in the first part of the survey, before showing the stimuli. This construct was assessed by four items, each measuring connectedness to a different group: *I feel strongly connected to…* (1) *people of the same age as me* (*M* = 4.88, *SD* = 1.50), (2) *older people* (*M* = 4.50, *SD* = 1.45), (3) *people with the same gender identity as me* (*M* = 5.31, *SD* = 1.32), and (4) *people with a different gender identity* (*M* = 4.47, *SD* = 1.42). These items were assessed on a 7-point scale ranging from (1) *completely disagree* to (7) *completely agree* and were included separately in the analyses.

After answering participants' perceptions on OHS for each vignette, participants' previous involvement in OHS was assessed. Two single-item measurements were included in the present study, both were rated on a 9-point scale ranging from (1) *never* to (9) *almost every hour* (based on: Kümpel & Unkel, 2023; Obermaier & Schmuck, 2022). First, frequency of being a pure bystander of OHS was measured with the question: *How often do you see user comments that contain negative stereotypes about other groups (groups based on e.g., gender, ethnicity, religion, sexual orientation, disability, …) you do not belong to online?* (*M* = 5.70, *SD* = 1.83). Second, frequency of being a vicarious victim of OHS was measured with the question: *How often do you see user comments that contain negative stereotypes about a group (group based on e.g., gender, ethnicity, religion, sexual orientation, disability, …) you belong to online?* (*M* = 5.10, *SD* = 2.10).

Participants

In total, 651 young adult social media users completed the questionnaire and passed two attention checks placed after each vignette between the items measuring participants' intention to perform prosocial bystander behavior. Sixteen participants were excluded due to speeding and two participants due to providing nonsense in one of the open question fields. The analytical sample consisted of 633 participants of which 50.7% (*n* = 321) identified as female, 47.9% (*n* = 303) as male, and 1.4% (*n* = 9) as non-binary, with age ranging from 18 to 25 (*M* = 21.81; *SD* = 2.34). Additionally, 97.3% (*n* = 616) participants had the Dutch nationality and 25.6% (*n* = 162) of participants hold a university degree. All participants used multiple types of social media at least once a month. Participants indicated for different applications how often they typically use these in their daily life: (1) *never*, (2) *less frequently than once a month*, (3) *once a month*, (4) *two to three times a month*, (5) *once a week*, (6) *several times a week*, (7) *once a day*, (8) *several times a day*, and (9) *(almost) hourly* (Kümpel & Unkel, 2023). Messenger apps were, on average, used most often (*multiple times per day*; *M* = 8.00, *SD* = 1.05), followed by social network sites (*M* = 7.30, *SD* = 1.72) and video platforms (*one time per day; M* = 6.80, *SD* = 1.51), and news sites/accounts (*multiple times per week*; *M* = 6.24, *SD* = 1.88), whereas microblog services were least often used (few times per month; *M* = 3.56, *SD* = 2.55).

The number of participants per condition is displayed in Appendix A. By means of Chi-Square tests and ANOVAs, conditions were compared for the socio-demographic characteristics gender identity, $\chi^2(18) = 10.65$, *p* = .909, *V* = .092 and age, *F*(9,623) = 1.22, *p* = .282, η^2 = .017, and social media use (microblog services: *F*(9,623) = 1.18, *p* = .304, η^2 = .017; social network sites: *F*(9,623) = .58, *p* = .811, η^2 = .008; video platforms: *F*(9,623) = 1.19, *p* = .302, η^2 = .017; messenger apps: *F*(9,623) = 0.54, *p* = .848, η^2 = .008; news sites/accounts: *F*(9,623) = 0.75, *p* = .655, η^2 = .011): no significant differences between conditions were found.

Participants were recruited through a commercial online access panel hosted by *Dynata*. All participants provided active informed consent. Data collection took place in December 2023 until February 2024. Participants were compensated by *Dynata*. The study was approved by the Ethics Committee of Tilburg University. To determine an appropriate sample size, a power analysis was calculated in the program G*Power v3.1.9.7 (Faul et al., 2007; Test Family: F-test; statistical test: analysis of variance (ANOVA) repeated measures, within-between interaction; Input parameters: effect size: .10, α-error: .05, power: .95, number of groups 10, number of measurements 2, correlation among repeated measures: .5), showing that 600 participants are needed. The data described in this article, except for the open-ended answers, are openly available in the Open Science Framework.

Data Analysis

Before analyzing the data, manipulation checks were conducted. First, participants were asked, at the end of the online survey, to indicate whether the perpetrators of the OHS reactions provided an online moral disengagement strategy to excuse their immoral behaviors. More than 9 out of ten participants (93.0%) correctly indicated that the perpetrators provided a statement to excuse their immoral behavior when there was indeed an online moral disengagement strategy statement included. Furthermore, participants indicated who were targeted by the two reactions they had seen. Almost all participants correctly indicated whether somebody of their gender identity ingroup or outgroup was targeted (95.9% correct answers). However, for OHS directed towards an age group, 77.9% correctly indicated whether somebody of their age ingroup or outgroup was targeted. Of those who failed (*N* = 140, 22.1 %) this manipulation check, 60.7% (*N* = 85) were in a condition in which they were pure bystander (older adults were targeted in the OHS) and 39.3% (*N* = 55) were in a condition in which they were vicarious victim (somebody from their own age group was attacked). Consequently, there was no clear sign whether one of these two roles was less clear/less well manipulated. A consultation of the literature indicated that there is some debate among experimentalists on how to proceed when a portion of the participants fails manipulation checks. The present study followed the suggestion from Mize and Manago (2022) to keep those who failed the manipulation checks as excluding those who failed could cause experimental conditions to become unbalanced, with an overrepresentation of certain types of participants in certain conditions (Aronow et al., 2019; Mize & Manago, 2022). To check whether the inclusion of these participants affected the results, the two main analyses of the present study (two Mixed ANOVAs) were also performed without those who failed the manipulation checks. This did not result in any changes regarding the acceptance/rejection of the hypotheses and, therefore, the presentation of the results in the results section is based on *N* = 633, including those who failed the manipulation checks.

Coding of the Open Questions

The answers to the two open questions in which participants were asked to label the OHS comments were cleaned and coded by the author. First, the answers were screened. Non-relevant responses (e.g., *I don't know*, age group: *N* = 275, gender identity: *N* = 172) and responses that referred to the participant's (dis)agreement with the OHS message were removed (e.g., *I disagree with this comment*; age group: *N* = 89, gender identity: *N* = 113). This resulted in the removal of answers of *N* = 364 (57.50%) participants for OHS targeting an age group and *N* = 285 (45.18%) participants for OHS targeting a certain gender identity. This means that the responses that were included in the analyses represented only a subset of the participants (OHS targeting an age group: *N* = 269, 42.5%; OHS targeting a gender identity: *N* = 348, 57.5%) and not the whole sample. Moreover, the included responses were also unevenly distributed among the ten different conditions (OHS targeting an age group: *N* ranging between 17 – 32; OHS targeting a gender identity: *N* ranging between 24 – 45), but in in each condition there were sufficient participants to execute the planned analyses. Next, the answers to one of the two open questions were open-coded. After this first round of coding, the codes were grouped and refined. This resulted in six categories: (1) OHS, (2) online aggression, (3) online incivility, (4) negative online reaction, (5) freedom of speech/neutral online reaction, (6) positive online reaction. The original coding of the first question was adapted, applying the final set of codes. This set was also used to code the second open question. Appendix B presents examples of participants' answers related to the six different categories and definitions used to categorize answers.

For further analysis, these six labels were dummy coded into two categories that provided an indication of the sentiments of the labels: (0) negative (containing the first four categories), (1) neutral to positive (containing the fifth and sixth category). This dummy variable was first used to perform a McNemar test, a repeated measures test for comparing frequency distributions. This test was used to investigate differences in labeling OHS as negative or neutral to positive between OHS targeting a certain age group and OHS targeting a certain gender identity. Furthermore, Pearson's Chi-Square tests were performed to statistically test differences in labeling (dummy variable) between OHS containing an online moral disengagement strategy statement (four types) or not, and differences in labeling (dummy variable) between pure bystanders and vicarious victims.

Main Analyses

Mixed ANOVAs (Mixed Analysis of Variance) were calculated to investigate differences in perceived offensiveness and intention to intervene with prosocial bystander behavior, accounting for both the effect of the between-subjects (moral excuses and the receiver's role) and within-subjects (repeated measures; OHS targeted on age and gender identity) factors, and of the control variables (previous exposure with OHS as a bystander and victim, and connectedness to the groups attacked).

Results

Attitudinal Component Towards OHS: Labeling

The first part of the present results section deals with the ways participants labeled the vignettes in the open questions. Participants' responses were recoded into a dummy variable that represented the sentiment of the label: (0) negative or (1) neutral to positive. As described earlier, first, a McNemar test was performed to test whether there were differences in labeling an OHS as negative or neutral to positive between OHS targeting a certain age group and OHS targeting a certain gender identity. Although this result cannot be linked to a specific hypothesis, it provides some information about potential differences in attitudes of bystanders towards OHS targeting a certain age group versus OHS targeting a certain gender identity (within factor). The test showed a significant difference, $\chi^2(1) = 12.89$, $p < .001$, $OR = 6.00$, indicating that OHS targeted on age was labeled by a larger portion of participants as neutral to positive (age: portion of participants that labeled OHS as neutral to positive: 21.9%) compared OHS targeted on gender (gender: portion of participants that labeled OHS as neutral to positive: 6.0%).

To compare differences in labeling between participants that were exposed to OHS containing an online moral disengagement strategy and participants that were exposed to OHS without an online moral disengagement strategy (H1a), two Pearson's Chi-Square tests were performed (crosstab with two variables: (1) online moral disengagement, consisting of five levels (four strategies and absence of a strategy) and (2) labels: negative versus neutral to positive), one for each targeted group characteristic (age and gender identity). Both tests (age: $\chi^2(4) = 1.62$, $p = .805$; $V = .078$; gender identity: $\chi^2(4) = .1.28$, $p = .866$; $V = .061$) indicated no significant differences in labeling OHS as negative versus neutral to positive between participants exposed to OHS containing one of the online moral disengagement strategies and participants exposed to OHS without an online moral disengagement strategy.

Two similar tests were performed to test whether OHS received by a pure bystander (OHS is directed towards an outgroup member) is labeled by a lower percentage as negative compared to when OHS is received by a vicarious victim (OHS is directed towards an ingroup member; H2a). Both tests (age: $\chi^2(1) = 0.93$, $p = .335$, $\varphi = -.059$; gender identity: $\chi^2(1) = 1.31$, $p = .253$, $\varphi = -.061$) indicated no significant differences in labeling OHS as negative versus neutral to positive when OHS is directed towards an outgroup member (bystander is a pure bystander) versus an ingroup member (bystander is a vicarious victim). As the labels provide an indication of one's attitude towards OHS, these results seem to suggest rejecting hypotheses H1a and H2a. These hypotheses predicted that the attitude towards OHS is less negative when an online moral disengagement strategy statement is included (H1a) in OHS and when OHS is directed towards (a member of) the outgroup (bystander's role = pure bystander; H2a). However, this preliminary conclusion should be interpreted with caution as (1) only a subset of the participants provided usable answers and (2) the categorization of the labels is based on the interpretation of the author.

Attitudinal Component Towards OHS: Perceived Offensiveness

To test the effects of online moral disengagement strategies statements and the receiver's role (pure bystander or vicarious victim) on perceived offensiveness (H1a & H2a), while controlling for previous experience with OHS as a bystander and victim and connectedness with the target groups, a Mixed ANOVA was calculated, accounting for the within factor (OHS targeted on two different groups, based on gender identity and age). Table 1 presents the mean scores for perceived offensiveness for each level of the two independent variables.

First, the model showed no main effect of the target group (within factor) on perceived offensiveness, $F(1,608) = 3.63$, $p = .057$, $\eta^2_p = .006$. Second, the model showed no main effects of online moral disengagement strategies, $F(4,608) = 0.65$, $p = .628$, $\eta^2_p = .004$ and the bystander's role (pure bystander or vicarious victim), $F(1,608) = 2.29$, $p = .131$, $\eta^2_p = .004$. Also no interaction effect between the bystander's role and online moral disengagement strategies on perceived offensiveness was found, $F(4,608) = 1.61$, $p = .170$, $\eta^2_p = .010$. However, the model did show a significant interaction effect of the target group (within factor) and the bystander's role (between factor; pure bystander or vicarious victim), $F(1,608) = 4.88$, $p = .028$, $\eta^2_p = .008$, indicating that there were significant differences between OHS directed towards an outgroup (pure bystander) and an ingroup (vicarious victim) for the perceived offensiveness of OHS, but only for OHS targeting gender identity ($B = -.53$, $SE = 0.23$, t = −2.31, $p = .022$, $\eta^2_p = .009$) and not OHS targeting age ($B = -.34$, $SE = 0.22$, $t = -1.582$, $p = .114$, $\eta^2_p = .004$). As predicted, OHS directed towards the ingroup was perceived as more offensive (M = 5.60, SD = 1.32) compared to OHS directed towards the outgroup ($M = 5.33$, $SD = 1.33$), but this was thus only true for OHS targeting gender identity. Regarding the control variables, no significant relations were found with perceived offensiveness. More precisely, both previous experience with OHS as a bystander, $F(1,608) = 3.36$, $p = .067$, $\eta^2_p = .005$ and as a vicarious victim, $F(1,608) = 0.17$, $p = .677$, $\eta^2_p = .000$, did not have an effect on perceived offensiveness. None of the included connectedness variables had an effect on perceived offensiveness: connectedness with same age individuals $F(1,608) = 1.90$, $p = .168$, $\eta^2_p = .003$; connectedness with older individuals $F(1,608) = 0.77$, $p = .381$, $\eta^2_p = .001$; connectedness with individuals with the same gender identity $F(1,608) = 2.51$, $p = .114$, $\eta^2_p = .004$; and connectedness with individuals with a different gender identity $F(1,608) = 0.40$, $p = .527$, $\eta^2_p = .001$.

Table 1. *Mean Scores for Perceived Offensiveness and Intention to Intervene With Prosocial Bystander Behavior.*

	Perceived offensiveness (age group targeted)	Perceived offensiveness (gender identity targeted)	Intention to perform prosocial bystander behavior (age group targeted)	Intention to perform prosocial bystander behavior (gender identity targeted)
	M (*SD*), *N*	*M* (*SD*), *N*	*M* (*SD*), *N*	*M* (*SD*), *N*
OHS towards outgroup	4.74 (1.22), 312	5.33 (1.33), 306	3.23 (1.64), 312	3.30 (1.65), 306
OHS towards ingroup	4.77 (1.27), 321	5.60 (1.32), 318	3.36 (1.56), 321	3.91 (1.70), 318
No online moral disengagement strategy statement present in OHS	4.84 (1.18), 136	5.55 (1.24), 135	3.43 (1.58), 136	3.74 (1.79), 135
Moral justification present in OHS	4.59 (1.37), 130	5.39 (1.37), 128	3.23 (1.66), 130	3.41 (1.74), 128
Diffusion of responsibility present in OHS	4.74 (1.14), 124	5.51 (1.33), 122	3.16 (1.54), 124	3.69 (1.62), 122
Distortion of consequences present in OHS	4.69 (1.21), 108	5.40 (1.38), 106	3.29 (1.55), 108	3.63 (1.50), 106
Attribution of blame present in OHS	4.92 (1.30), 135	5.47 (1.36), 133	3.36 (1.67), 135	3.60 (1.79), 133

Note. Both perceived offensiveness and intention to perform prosocial bystander behavior were rated on a 7-point scale ranging from (1) *completely disagree* to (7) *completely agree*, with higher scores presenting stronger perceived offensiveness and higher intention to perform prosocial bystander behavior.

Behavioral Component Towards OHS: Intention to Intervene With Prosocial Bystander Behavior

The second Mixed ANOVA tested the effects on intention to intervene with prosocial bystander behavior (H1b & H2b; the same independent and controlling variables were used as in the previous Mixed ANOVA). Table 1 also presents the mean scores for intention to intervene with prosocial bystander behavior for each level of the two independent variables.

First, the model did not show a significant main effect of the target group (within factor) on intention to intervene, *F*(1,608) = 0.35, *p* = .556, η^2_p = .001. Second, no significant main effect was found for online moral disengagement strategies on the intention to intervene, *F*(4,608) = 0.52, *p* = .723, η^2_p = .003. Third, the model showed a significant main effect of the bystander's role, *F*(1,608) = 8.62, *p* = .003, η^2_p = .014, on perceived intention to intervene. The mean intention to intervene with prosocial bystander behavior was higher when OHS was directed towards an ingroup (bystander = vicarious victim: *M* = 3.61, *SD* = 0.08), compared to an outgroup (bystander = pure bystander: *M* = 3.30, *SD* = 0.08). A significant interaction effect between the bystander's role (pure bystander versus vicarious victim) and the target group on intention to intervene, *F*(1,608) = 17.99, *p* < .001, η^2_p = .029, indicated that the difference between OHS directed towards the ingroup and OHS directed towards the outgroup on intention to intervene was larger for OHS targeting gender identity (outgroup: *M* = 3.30, *SD* = 1.65; ingroup: *M* = 3.91, *SD* = 1.70) than for OHS targeting age (outgroup: *M* = 3.23, *SD* = 1.64; ingroup: *M* = 3.36, *SD* = 1.56). Fourth, no interaction effect between online moral disengagement statement strategies and the bystander's role on intention to intervene was found, *F*(4,608) = 2.18, *p* = .07, η^2_p = .014.

Regarding the individual traits control variables, both previous experience with OHS as a bystander, *F*(1,608) = 12.41, *p* < .001, η^2_p = .020 and as a vicarious victim, *F*(1,608) = 21.54, *p* < .001, η^2_p = .034, had an effect on intention to intervene with prosocial bystander behavior. The more frequent participants were exposed to OHS directed at an outgroup (bystander) in the past, the lower the intention to intervene with prosocial bystander behavior (age: *B* = −.15 , *SE* = 0.04, *t* = −3.56, *p* < .001, η^2_p = .020); gender identity: *B* = −.12 , *SE* = 0.04, *t* = −2.68, *p* = .008, η^2_p = .012), whereas the more frequent participants were exposed to OHS directed towards an ingroup (vicarious victim) in the past, the higher the intention to intervene with prosocial bystander behavior (age: *B* = .16, *SE* = 0.04, *t* = 4.47, *p* < .001, η^2_p = .032); gender identity: *B* = .14 , *SE* = 0.04, *t* = 3.74, *p* < .001, η^2_p = .023). As the coefficients indicate, this was true for both group characteristics. Finally, connectedness with older individuals had an effect on intention to intervene, *F*(1,608) = 5.94, *p* = .015, η^2_p = .010. The stronger participants felt connected to older individuals, the higher their intention to intervene with prosocial bystander behavior when witnessing OHS targeting an age group (age: *B* = .14, *SE* = 0.05, *t* = 3.03, *p* = .003, η^2_p = .015). The coefficients indicated that this was only true for OHS targeting an age group, not for OHS targeting a gender identity group (gender identity: *B* = .07, *SE* = 0.05, *t* = 1.32, *p* = .187, η^2_p = .003). Connectedness to people of the same age did not have an effect on the intention to intervene with prosocial behavior, *F*(1,608) = 0.54, *p* = .464, η^2_p = .001. Moreover, also connectedness to people with the same gender identity, *F*(1,608) = 0.47, *p* = .492, η^2_p = .001 and connectedness to people with a different gender identity, *F*(1,608) = 3.24, *p* = .072, η^2_p = .005, did not have an effect on the intention to intervene with prosocial bystander behavior when witnessing OHS targeting a group based on gender identity.

Discussion

The presented study aimed to predict bystanders' attitudinal and behavioral perceptions of OHS, based on the presence or absence of online moral disengagement strategies statements in OHS messages (content-related characteristic) and the bystander's role (pure bystander or vicarious victim; contextual determinant), while controlling for previous experience with OHS (personal characteristic) and connectedness with the target group (personal characteristic).

First, the results indicated, for both the attitudinal and behavioral components of bystander's perceptions on OHS, no differences between the four different online moral disengagement strategies and the absence of such a statement in OHS, which means H1a and H1b should be rejected. In other words, it seems that the presence of a moral excuse has no influence on (1) how bystanders label the type of speech they have seen (as negative or rather neutral to positive), (2) bystanders' perceived offensiveness of OHS, and (3) bystanders' intention to intervene with prosocial bystander behavior. These findings contradict with the results of Wilhelm et al. (2020)

and expectations based on theoretical frameworks such as neutralization theory (Sykes & Matza, 1957) and moral disengagement (Bandura, 1986, 1999). In the present study, these statements probably did not ‘mask’ (enough) the intentions of the perpetrator, and/or did not make OHS appear more morally acceptable. The perceived offensiveness of the OHS messages was potentially of such a high level rendering any justification as ineffective. This was supported by the mean scores, being (far) above the midpoint of the answer option scale, for both OHS without online moral disengagement strategies and OHS with online moral disengagement strategies. An alternative explanation might be that without expressing these justifications, bystander could still imagine that the perpetrator had motivations in line with these justifications to post OHS, as they have seen these justifications before in OHS. Future research might want to dive deeper into perceptions of bystanders on motivations of perpetrators to post OHS.

Second, the present study investigated the bystander’s role, or, in other words, whether pure bystanders (OHS is directed towards an outgroup (member)) perceive OHS differently compared to vicarious victims (OHS is directed towards an ingroup member). Investigating this contextual characteristic was driven by a recent call in the literature to investigate vicarious victimization as a separate role (Stahel & Baier, 2023). Based on theories such as social identity, it was expected that bystanders who are vicarious victims perceive OHS as more negative (attitudinal component) and have a higher intention to intervene with prosocial behavior (behavioral component). The data provided evidence for these expectations (H2a).

More precisely, for perceived offensiveness, vicarious victims perceived OHS as more offensive compared to pure bystanders when witnessing OHS based on gender identity. However, we could not replicate this finding for OHS based on age. These results might stimulate future research to dive deeper into potential differences between target groups. For instance, Obermaier et al. (2023) showed that homophobic hate speech was perceived to be less uncivil than OHS against women. Furthermore, we also want to note that the present study controlled for the potential connectedness with these groups (as suggested by previous research, e.g., Kümpel & Unkel, 2023; Papcunová et al., 2023; Rudnicki et al., 2022), but also these variables did not explain the presence of the significant difference between pure bystanders and vicarious victims when OHS is targeting a gender identity group and the absence of this difference when OHS is targeting an age group. A potential explanation between the differential results for gender-related and age-related OHS might be explained by the presentation order of the vignettes. All participants first saw a vignette targeting a specific age group and next a vignette targeting a specific gender identity group. Exposure to the first vignette might have influenced reactions on the second vignette. Finally, it should be mentioned that the labeling question (open question) did not show differences between pure bystanders and vicarious victims in labeling the speech as negative or rather neutral to positive.

Regarding the behavioral component of bystanders’ perceptions, the results indicated that, as hypothesized by H2b, vicarious victims had a significant higher intention to intervene with prosocial bystander behavior than pure bystanders. This was true for both OHS targeting a certain age group and OHS targeting a certain gender identity. However, the results did show that this difference in intention between vicarious victims and pure bystanders was larger when exposed to OHS targeting gender identity. Again, this asks for further exploration regarding differences in perceptions of bystanders towards OHS targeting different groups, but also raises questions about potential effects of the presentation order of the vignettes. It should be noted that the smaller difference between pure bystanders and vicarious victims in intention to intervene when witnessing OHS targeting a certain age group might be explained by a significant association between intention to intervene and one of the control variables, namely connectedness with older adults. The latter was positively associated with intention to intervene with prosocial bystander behavior when witnessing OHS targeted on age: the more strongly participants felt connected to older adults, the higher their intention to intervene.

While investigating differences in perceived offensiveness and intention to intervene with prosocial bystander behavior, the present study did not only account for connectedness, but also for two individual factors, namely previous exposure to OHS as a bystander and as a vicarious victim. Previous experience (frequency) with OHS as a bystander and vicarious victim did not have an effect on perceived offensiveness. In other words, being exposed to OHS frequently in the past was not associated with finding it more (as shown, for instance, by: Costello et al., 2019), or less offensive (as shown, for instance, by: Schäfer et al., 2022; Schmid et al., 2024), the latter would have been an indication for desensitization to and normalization of OHS. For the behavioral component of bystanders’ perceptions, however, significant associations were found with both previous experience with OHS as a bystander and as a vicarious victim. The more frequent participants were exposed to OHS as a pure bystander in the past (in other words, being exposed to OHS directed at an outgroup), the lower the indicated intention to intervene with prosocial bystander behavior when being exposed to OHS in the present study. This finding supports processes of desensitization to and normalization of OHS, as described by previous studies (e.g., Schäfer et al., 2022; Schmid et al., 2024). In contrast, frequent previous exposure to OHS as a vicarious victim was associated with higher intention to intervene with prosocial bystander behavior when being exposed to OHS in the present study (in line with: Costello et al., 2023; Mohseni, 2023; Ping et al., 2025). The latter might indicate that frequent vicarious victims feel more strongly personally responsible for intervening when witnessing OHS, even when OHS is not directed to their ingroup.

Finally, the present study also gave some insights, based on qualitative data resulting from open questions in the survey, on different labels or terms young adult social media users use for referring to speech that can be defined as OHS, according to the (academic) definition that was used in the present study (Hawdon et al., 2017; Johnson et al., 2019). These qualitative data were used to explore attitudinal differences, next to the data that resulted from the quantitative attitude measurement based on closed questions (perceived offensiveness). However, it is worth noting that there was a large variety of terms that were used by the participants to label the OHS comments (see Appendix B). This variety implies that people do use different terms for behavior that is defined in the academic literature as OHS and stresses the need to explain clearly what is meant exactly when using the term OHS in research, but, for instance, also in intervention and prevention campaigns. Moreover, even in the academic literature different terms are used to refer to speech that is defined in the present study as OHS, see for instance Frischlich (2023) or Kümpel and Unkel (2023). Clarifications about what is meant exactly by the terms that are being used seem crucial. It should be noted that a large number of answers to the open questions (about half of the answers in total) needed to be removed and that the included answers were unevenly distributed across the conditions. Therefore, the results related to labeling should be interpreted with caution. Future research that wants to include an open question to grasp bystanders’ attitudes via sentiments expressed in open answers might want to consider giving extra instructions to the bystander, for instance, indicating that answers should not involve bystanders’ agreement or disagreement with the opinion expressed in the OHS message. Future research might also want to double code (part of) the dataset and check intercoder reliability.

Based on the results, practical and theoretical implications can be formulated. The results indicate that bystanders’ perceptions differ when an ingroup is attacked compared to an outgroup. This could imply that positive bystander behavior will be more often performed when groups that are more present online are targeted, and, on the other hand, victims of OHS will be less likely supported if they differ from those groups that are more present online. Researchers have advocated to promote critical media literacy, from an early age on, to counter OHS (see, for instance, Bruschi et al., 2023). Education on critical media literacy should make social media users able to “observe, analyze, evaluate, create and fully – but -safely – participate to the digital life” (Bruschi et al., 2023, p. 8), which does not only mean taking responsibility for your own behaviors, but also critically consume content of others and intervening when offensive speech is spread (e.g., Naderer et al., 2023; Pukallus & Arthur, 2024).

Implications can also be formulated regarding the lack of evidence for the influence of the use of (different) online moral disengagement strategies statements on bystanders’ perceptions. The results seem to imply that when perpetrators share these beliefs, others will not (automatically) accept and refrain from taking action. This contradicts with theoretical frameworks such as neutralization theory (Sykes & Matza, 1957) and moral disengagement (Bandura, 1986, 1999). However, there might be specific conditions in which bystanders will (automatically) accept, for instance, when they have limited cognitive capacity due to (an abundance of) peripheral cues, such as when the message is ‘hidden’ in a (video) post containing a high amount of visual and auditory cues, e.g., the presence of animations, humor, sound effects, and so on (Petty & Cacioppo, 1986; Schmid, 2023). In such conditions, bystanders might make judgements based on heuristics instead of evaluating the speech critically.

The present study has shortcomings. First, the used stimuli required some imagination from the side of the bystander as the vignettes contained written text, but, for instance, did not contain images representing a certain media platform or profiling the sender of the OHS. It was also not specified on which platform the incidents described took place, nor who exactly was posting the reaction. In addition, information about other contextual and content-related factors that have been found to influence bystanders' perceptions were not provided, such as the number of bystanders and prior reactions of bystanders (Costello et al., 2023; Leonhard et al., 2018). (The lack of all) these elements influenced the ecological validity of the experiment. Future studies could consider exposing participants to 'real life' situations that are collected through data donations from other users (Ohme et al., 2023). Furthermore, instead of behavioral intentions to intervene, future research might also want to include measurements of actual behavior, for instance by allowing participants to immediately react by writing a counterspeech message or a supporting message directed to the victim. Finally, differential results were found for age-related and gender-related hate speech. Potential explanations were provided, one being related to the study design, in which the presentation order of the vignettes was not randomized. The absence of randomization of the presentation order of the vignettes could be considered as a limitation of the present study, as exposure to and one's reaction on the first vignette could have influenced one's reaction to the second vignette, whereas this was not true for the first vignette. The lack of randomization could potentially also explain why more participants failed the manipulation checks for the age group that was targeted in the first OHS that they had seen, it is likely that some participants could not recall this group after being exposed to another OHS targeting a certain gender identity.

To conclude, the goal of the present study was to elaborate our understanding of perceptions of bystanders on OHS by investigating content-related, contextual, and personal characteristics. By means of an experimental design, the present study showed the importance of the bystander's role for forming bystanders' perceptions on OHS. The main takeaway of the present study is that bystanders who are exposed to OHS targeting an individual with whom the bystander shares the targeted group characteristic perceive OHS as more offensive and have a higher intention to intervene with prosocial bystander behavior, compared to bystanders who do not share the group characteristic under attack in the OHS. This finding seems to imply that victims of OHS will be less likely supported if they differ from those groups that are more present online.

Conflict of Interest

The author has no conflicts of interest to declare.

Use of AI Services

The author declares to have used AI services, specifically MS Word and DeepL, for grammar correction and minor style refinements. The author carefully reviewed all suggestions from these services to ensure the original meaning and factual accuracy were preserved.

Author‘s Contribution

This study was devised and conducted by **Sara Pabian.**

Acknowledgement

This research is funded by the Dutch Research Council (NWO; project number 406.XS.01.027).

References

Aronow, P. M., Baron, J., & Pinson, L. A (2019). A note on dropping experimental subjects who fail a manipulation check. *Political Analysis, 27*(4), 572–589. https://doi.org/10.1017/pan.2019.5

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice-Hall.

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

Batson, C. D. (1991). *The altruism question: Toward a social-psychological answer*. Lawrence Erlbaum. https://doi.org/10.4324/9781315808048

Bormann, M., Tranow, U., Vowe, G., & Ziegele, M. (2022). Incivility as a violation of communication norms: A typology based on normative expectations toward political communication. *Communication Theory, 32*(3), 332–362. https://doi.org/10.1093/ct/qtab018

Bruschi, B., Repetto, M., & Talarico, M. (2023). A framework on media-educational initiatives to contrast online hate speech. *Q-Times Webmagazine, 2*(1), 7–16. https://iris.unito.it/handle/2318/1894026

Cardwell, S. M., & Copes, H. (2021). Neutralization. In B. van Rooij & D. D. Sokol (Eds.), *The Cambridge handbook of compliance* (pp. 451–464). Cambridge University Press. https://doi.org/10.1017/9781108759458.031

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

Catelli, R., Pelosi, S., Comito, C., Pizzuti, C., & Esposito, M. (2023). Lexicon-based sentiment analysis to detect opinions and attitude towards COVID-19 vaccines on Twitter in Italy. *Computers in Biology and Medicine, 158*, Article 106876. https://doi.org/10.1016/j.compbiomed.2023.106876

Costello, M., Hawdon, J., Bernatzky, C., & Mendes, K. (2019). Social group identity and perceptions of online hate. *Sociological Inquiry, 89*(3), 427–452. https://doi.org/10.1111/soin.12274

Costello, M., Hawdon, J., Reichelmann, A. V., Oksanen, A., Blaya, C., Llorent, V. J., Räsänen, P., & Zych, I. (2023). Defending others online: The influence of observing formal and informal social control on one's willingness to defend cyberhate victims. *International Journal of Environmental Research and Public Health, 20*(15), Article 6506. https://doi.org/10.3390/ijerph20156506

D'Errico, F., & Paciello, M. (2018). Online moral disengagement and hostile emotions in discussions on hosting immigrants. *Internet Research, 28*(5), 1313–1335. https://doi.org/10.1108/IntR-03-2017-0119

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

Domínguez-Hernández, F., Bonell, L., & Martínez-González, A. (2018). A systematic literature review of factors that moderate bystanders' actions in cyberbullying. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, *12*(4), Article 1. https://doi.org/10.5817/CP2018-4-1

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/BF03193146

Faulkner, N., & Bliuc, A.-M. (2016). 'It's okay to be racist': Moral disengagement in online discussions of racist incidents in Australia. *Ethnic and Racial Studies, 39*(14), 2545–2563. https://doi.org/10.1080/01419870.2016.1171370

Fraser, K. C., Kiritchenko, S., & Nejadgholi, I. (2022). Extracting age-related stereotypes from social media texts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 3183–3194). European Language Resources Association. https://aclanthology.org/2022.lrec-1.341/

Frischlich, L. (2023). Hate and harm. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 165–183). https://doi.org/10.48541/dcr.v12.10

Greer, C., & Jewkes, Y. (2005). Extremes of otherness: Media images of social exclusion. *Social Justice, 32*(1), 20–31. https://www.jstor.org/stable/pdf/29768287.pdf

Grigg, D. W. (2010). Cyber-aggression: Definition and concept of cyberbullying. *Australian Journal of Guidance and Counselling, 20*(2), 143–156. https://doi.org/10.1375/ajgc.20.2.143

Gullette, M. M. (2024). Are older people still human? On ageist humor. In V. B., Lipscomb & A. Swinnen (Eds.), *The Palgrave handbook of literature and aging* (pp. 569–589). Palgrave Macmillan. https://doi.org/10.1007/978-3-031-50917-9_29

Hawdon, J., Oksanen, A., & Räsänen, P. (2017). Exposure to online hate in four nations: A cross-national consideration. *Deviant Behavior, 38*(3), 254–266. https://doi.org/10.1080/01639625.2016.1196985

Henry, S. (2009). *Social deviance*. Polity Press.

Henson, B., Fisher, B. S., & Reyns, B. W. (2020). There is virtually no excuse: The frequency and predictors of college students' bystander intervention behaviors directed at online victimization. *Violence Against Women, 26*(5), 505–527. https://doi.org/10.1177/1077801219835050

Hogg, M. A., & Abrams, D. (1988). *Social identifications: A social psychology of intergroup relations and group processes*. Routledge. https://doi.org/10.4324/9780203135457

Johnson, N. F., Leahy, R., Restrepo, N. J., Velasquez, N., Zheng, M., Manrique, P., Devkota, P., & Wuchty, S. (2019). Hidden resilience and adaptive dynamics of the global online hate ecology. *Nature, 573*(7773), 261–265. https://doi.org/10.1038/s41586-019-1494-7

Kenski, K., Coe, K., & Rains, S. A. (2020). Perceptions of uncivil discourse online: An examination of types and predictors. *Communication Research, 47*(6), 795–814. https://doi.org/10.1177/0093650217699933

Kümpel, A. S., & Rieger, D. (2019). *Wandel der Sprach- und Debattenkultur in sozialen Online-Medien: ein Literaturüberblick zu Ursachen und Wirkungen von inziviler Kommunikation* [Changes in language and debate culture in online social media: A literature review on the causes and effects of uncivil communication]. Konrad-Adenauer-Stiftung e. V. https://doi.org/10.5282/ubm/epub.68880

Kümpel, A. S., & Unkel, J. (2023). Differential perceptions of and reactions to incivil and intolerant user comments. *Journal of Computer-Mediated Communication, 28*(4), Article zmad018. https://doi.org/10.1093/jcmc/zmad018

Kunst, M., Porten-Cheé, P., Emmer, M., & Eilders, C. (2021). Do "good citizens" fight hate speech online? Effects of solidarity citizenship norms on user responses to hate comments. *Journal of Information Technology & Politics, 18*(3), 258–273. https://doi.org/10.1080/19331681.2020.1871149

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

Lewis, M., & Lupyan, G. (2020). Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature Human Behaviour, 4*, 1021–1028. https://doi.org/10.1038/s41562-020-0918-6

Macaulay, P. J. R., Betts, L. R., Stiller, K., & Kellezi, B. (2022). Bystander responses to cyberbullying: The role of perceived severity, publicity, anonymity, type of cyberbullying, and victim response. C*omputers in Human Behavior, 131*, Article 107238. https://doi.org/10.1016/j.chb.2022.107238

Maftei, A., Holman, A.-C., & Merlici, I.-A. (2022). Using fake news as means of cyber-bullying: The link with compulsive internet use and online moral disengagement. *Computers in Human Behavior, 127*, Article 107032. https://doi.org/10.1016/j.chb.2021.107032

Marín-López, I., Zych, I., Ortega-Ruiz, R., Monks, C. P., & Llorent, V. J. (2020). Empathy online and moral disengagement through technology as longitudinal predictors of cyberbullying victimization and perpetration. *Children and Youth Services Review, 116*, Article 105144. http://doi.org/10.1016/j.childyouth.2020.105144

Matamoros-Fernández, A., & Farkas, J. (2021). Racism, hate speech, and social media: A systematic review and critique. *Television & New Media, 22*(2), 205–224. https://doi.org/10.1177/1527476420982230

Mize, T. D., & Manago, B. (2022). The past, present, and future of experimental methods in the social sciences. *Social Science Research, 108*, Article 102799. https://doi.org/10.1016/j.ssresearch.2022.102799

Mohseni, M. R. (2023). Motives of online hate speech: Results from a quota sample online survey. *Cyberpsychology, Behavior, and Social Networking, 26*(7), 499–506. https://doi.org/10.1089/cyber.2022.0188

Naderer, B., Wendt, R., Bachl, M., & Rieger, D. (2023). Understanding the role of participatory-moral abilities, motivation, and behavior in European adolescents' responses to online hate. *New Media & Society, 27*(3), 1774–1794. https://doi.org/10.1177/14614448231203617

Nocera, T. R., Dahlen, E. R., Poor, A., Strowd, J., Dortch, A., & Van Overloop, E. C. (2022). Moral disengagement mechanisms predict cyber aggression among emerging adults. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace, 16*(1), Article 6. https://doi.org/10.5817/CP2022-1-6

Nurhadiyanto, L., & Octaviani, M. Y. (2021). Analysis of neutralization theory in hatespeech against Kekeyi Young Lex and Nissa Sabyan on Instagram. *ICCD, 3*(1), 202–207. https://doi.org/10.33068/iccd.Vol3.Iss1.338

Obermaier, M. (2024). 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

Obermaier, M., & Schmuck, D. (2022). Youths as targets: Factors of OHS victimization among adolescents and young adults. *Journal of Computer-mediated Communication, 27*(4), Article zmac012. https://doi.org/10.1093/jcmc/zmac012

Obermaier, M., Schmid, U. K., & Rieger, D. (2023). Too civil to care? How online hate speech against different social groups affect bystander intervention. *European Journal of Criminology, 20*(3), 817–833. https://doi.org/10.1177/14773708231156328

Ohme, J., & Mothes, C. (2020). What affects first- and second-level selective exposure to journalistic news? A social media online experiment. *Journalism Studies, 21*(9), 1220–1242. https://doi.org/10.1080/1461670X.2020.1735490

Ohme, J., Araujo, T., Boeschoten, L., Freelon, D., Ram, N., Reeves, B. B., & Robinson, T. N. (2023). Digital trace data collection for social media effects research: APIs, data donation, and (screen) tracking. C*ommunication Methods and Measures, 18*(2), 124–141. https://doi.org/10.1080/19312458.2023.2181319

Paasch-Colberg, S., Strippel, C., Trebbe, J., & Emmer, M. (2021). From insult to hate speech: Mapping offensive language in German user comments on immigration. *Media and Communication, 9*(1), 171–180. https://doi.org/10.17645/mac.v9i1.3399

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

Pabian, S., Vandebosch, H., Poels, K., Van Cleemput, K., & Bastiansens, 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

Paciello, M., D'Errico, F., & Saleri, G. (2019). Moral struggles in social media discussion: The case of sexist aggression. In B. N. De Carolis, F. D'Errico, & V. Rossano (Eds.), *Proceedings of the workshop socio-affective technologies: An interdisciplinary approach co-located with IEEE SMC 2019* (pp.13–16). CEUR workshop proceedings. https://ceur-ws.org/Vol-2474/shortpaper3.pdf

Palladino, B. E., Menesini, E., Nocentini, A., Luik, P., Naruskov, K., Ucanok, Z., Dogan, A., Schultze-Krumbholz, A., Hess, M., & Scheithauer, H. (2017). Perceived severity of cyberbullying: Differences and similarities across four countries. *Frontiers in Psychology, 8*, Article 1524. https://doi.org/10.3389/fpsyg.2017.01524

Papcunová, J., Martončik, M., Fedáková, D., Kentoš, M., & Adamkovič, M. (2023). Perception of hate speech by the public and experts: Insights into predictors of the perceived hate speech towards migrants. *Cyberpsychology, Behavior, and Social Networking, 26*(7), 489–498. https://doi.org/10.1089/cyber.2022.0191

Paschalides, D. Stephanidis, D., Andreou, A., Orphanou, K., Pallis, G., Dikaiakos, M. D., & Markatos, E. (2020). Mandola: A big-data processing and visualization platform for monitoring and detecting OHS. *ACM Transactions on Internet Technology, 20*(2), Article 11. https://doi.org/10.1145/3371276

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology, 19*, 123–205. https://doi.org/10.1016/s0065-2601(08)60214-2

Ping, K., Kumar, A., Ding, X., & Rho, E. (2025). Behind the counter: Exploring the motivations and barriers of online counterspeech writing. *ACM Transactions on Computer-Human Interaction*. https://doi.org/10.1145/3745769

Pukallus, S., & Arthur, C. (2024). Combating hate speech on social media: Applying targeted regulation, developing civil-communicative skills and utilising local evidence-based anti-hate speech interventions. *Journalism and Media, 5*(2), 467–484. https://doi.org/10.3390/journalmedia5020031

Reichelmann, A., Hawdon, J., Costello, M., Ryan, J., Blaya, C., Llorent, V., Oksanen, A., Räsänen, P., & Zych, I. (2021). Hate knows no boundaries: Online hate in six nations. *Deviant Behavior, 42*(9), 1100–1111. https://doi.org/10.1080/01639625.2020.1722337

Ribeaud, D., & Eisner, M. (2010). Are moral disengagement, neutralization techniques, and self-serving cognitive distortions the same? Developing a unified scale of moral neutralization of aggression. *International Journal of Conflict and Violence, 4*(2), 298–315. https://doi.org/10.4119/ijcv-2833

Rieger, D., Kümpel, A. S., Wich, M., Kiening, T., & Groh, G. (2021). Assessing the extent and types of hate speech in fringe communities: A case study of alt-right communities on 8chan, 4chan, and Reddit. *Social Media + Society, 7*(4). https://doi.org/10.1177/20563051211052906

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

Runions, K. C., & Bak, M. (2015). Online moral disengagement, cyberbullying, and cyber-aggression. *Cyberpsychology, Behavior, and Social Networking, 18*(7), 400–405. https://doi.org/10.1089/cyber.2014.0670

Schäfer, S., Sülflow, M., & Reiners, L. (2022). Hate speech as an indicator for the state of the society. Effects of hateful user comments on perceived social dynamics. *Journal of Media Psychology, 34*(1), 3–15. https://doi.org/10.1027/1864-1105/a000294

Schmid, U. K. (2023). Humorous hate speech on social media: A mixed-methods investigation of users' perceptions and processing of hateful memes. *New Media & Society, 27*(3), 1588–1606. https://doi.org/10.1177/14614448231198169

Schmid, U. K., Kümpel, A. S., & Rieger, D. (2024). How social media users perceive different forms of OHS: A qualitative multi-method study. *New Media & Society, 26*(5), 2614–2632. https://doi.org/10.1177/14614448221091185

Skitka, L. J., Hanson, B. E., & Wisneski, D. C. (2017). Utopian hopes or dystopian fears? Exploring the motivational underpinnings of moralized political engagement. *Personality and Social Psychology Bulletin, 43*(2), 177–190. https://doi.org/10.1177/0146167216678858

Stahel, L., & Baier, D. (2023). Digital hate speech experiences across age groups and their impact on well-being: A nationally representative survey in Switzerland. *Cyberpsychology, Behavior, and Social Networking, 26*(7), 519–526. https://doi.org/10.1089/cyber.2022.0185

Stets, J. E., & Burke, P. J. (2000). Identity theory and social identity theory. *Social Psychology Quarterly, 63*(3), 224–237. https://doi.org/10.2307/2695870

Stürmer, S., & Snyder, M. (2009). Helping "us" versus "them." In S. Stürmer & M. Snyder (Eds.), *The psychology of prosocial behavior: Group processes, intergroup relations, and helping* (pp. 33–58). Wiley-Blackwell. https://doi.org/10.1002/9781444307948.ch2

Sykes, G. M., & Matza, F. (1957). Techniques of neutralization: A theory of delinquency. *American Sociological Review, 22*(6), 664–670. https://doi.org/10.2307/2089195

Tajfel, H. (1974). Social identity and intergroup behaviour. *Social Science Information, 13*(2), 65–93. https://doi.org/10.1177/053901847401300204

Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. S. (1987). *Rediscovering the social group: A self-categorization theory*. Blackwell.

Wilhelm, C., Joeckel, S., & Ziegler, I. (2020). Reporting hate comments: Investigating the effects of deviance characteristics, neutralization strategies, and users' moral orientation. *Communication Research, 47*(6), 921–944. https://doi.org/10.1177/0093650219855330

Appendices

Appendix A

Table A1. *Stimuli Materials.*

Condition	Explanation	Vignette A	Vignette B
1. No OMD strategy – Bystander *n* = 66, 10.4%	OHS containing a negative stereotype of individuals (a) belonging to a different age cohort and (b) with a different gender identity, compared to the participant. No OMD strategy present.	1a) Imagine seeing a news article on a social media platform with the following headline: “65-year-old appointed as new CEO of largest government company in the Netherlands". Other news readers have posted a comment. Someone posted the following comment: *I really don't understand why they let the whole thing be run by an old person who will be buried in a few years and who I'm sure will put in less effort than a young person. Waste of time and money to appoint this person!*	1b) Imagine seeing a news article on a social media platform with the following headline: “Male driver hits child”. Other news readers have posted a comment. Someone posted the following comment: I really don’t understand why male drivers are not followed up. They are aggressive and make more often traffic violations! Males should take an annual driving test. Failed: turn in your driver's license! A traffic violation while driving? Turn it in! Or (depending on the participants own gender identity): Imagine seeing a news article on a social media platform with the following headline: “Female driver hits child”. Other news readers have posted a comment. Someone posted the following comment: I really don’t understand why female drivers are not followed up. They are insecure in traffic, and they cannot remember the traffic rules! Females should take an annual driving test. Failed: turn in your driver's license! A traffic violation while driving? Turn it in!
2. No OMD strategy – Vicarious victim *n* = 70, 11.1%	OHS containing a negative stereotype of individuals (a) belonging to the same age cohort and (b) with the same gender identity, as the participant. No OMD strategy present.	2a) Imagine seeing a news article on a social media platform with the following headline: “22-year-old appointed as new CEO of largest government company in the Netherlands". Other news readers have posted a comment. Someone posted the following comment: *I really don't understand why they let the whole thing be run by a young person who has no knowledge and who surely will be bored out of the job in a few years. Waste of time and money to appoint this person!*	See 1b
3. OMD strategy: Moral justification – Bystander *n* = 64, 10.1%	OHS containing a negative stereotype of individuals (a) belonging to a different age cohort and (b) with a different gender identity, compared to the participant. Moral justification claim present, protecting the commenter’s own group.	*1a + I’m posting this comment online because I feel that one should also give a somewhat younger person opportunities.*	1b + *I’m posting this comment online because I want to protect myself and important others.*
4. OMD strategy: Moral justification – Vicarious victim *n* = 66, 10.4%	OHS containing a negative stereotype of individuals (a) belonging to the same age cohort and (b) with the same gender identity, as the participant. Moral justification claim present, protecting the commenter’s own group.	*2a +I’m posting this comment online because I feel that one should also give a slightly older person opportunities.*	1b + *I’m posting this comment online because I want to protect myself and important others.*

5. OMD strategy: Diffusion of responsibility – Bystander *n* = 58, 9.2%	OHS containing a negative stereotype of individuals (a) belonging to a different age cohort and (b) with a different gender identity, compared to the participant. Diffusion of responsibility claim present, referring to others performing the same behavior.	1a + *I'm posting this comment because I see a lot of others are doing this online as well.*	1b + *I'm posting this comment because I see a lot of others are doing this online as well.*
6. OMD strategy: Diffusion of responsibility – Vicarious victim *n* = 66, 10.4%	OHS containing a negative stereotype of individuals (a) belonging to the same age cohort and (b) with the same gender identity, as the participant. Diffusion of responsibility claim present, referring to others performing the same behavior.	2a + *I'm posting this comment because I see a lot of others are doing this online as well.*	1b + *I'm posting this comment because I see a lot of others are doing this online as well.*
7. OMD strategy: Distortion of consequences – Bystander *n* = 55, 8.7%	OHS containing a negative stereotype of individuals (a) belonging to a different age cohort and (b) with a different gender identity, compared to the participant. Distortion of consequences claim present, referring to the belief that sharing your opinion online is harmless.	1a + *I'm posting this comment because it doesn't hurt anyway if I share my opinion online.*	1b + *I'm posting this comment because it doesn't hurt anyway if I share my opinion online.*
8. OMD strategy: Distortion of consequences – Vicarious victim *n* = 53, 8.4%	OHS containing a negative stereotype of individuals (a) belonging to the same age cohort and (b) with the same gender identity, as the participant. Distortion of consequences claim present, referring to the belief that sharing your opinion online is harmless.	2a + *I'm posting this comment because it doesn't hurt anyway if I share my opinion online.*	1b + *I'm posting this comment because it doesn't hurt anyway if I share my opinion online.*
9. OMD strategy: Attribution of blame – Bystander *n* = 69, 10.9%	OHS containing a negative stereotype of individuals (a) belonging to a different age cohort and (b) with a different gender identity, compared to the participant. Attribution of blame claim present, blaming the group that is targeted (it is their own fault that they are targeted).	1a + *I'm posting this comment because older people have themselves to blame for being ridiculed online.*	1b + *I'm posting this comment because females have themselves to blame for being ridiculed online.* Or (depending on the participants own gender identity): 1b + *I'm posting this comment because males have themselves to blame for being ridiculed online.*
10. OMD strategy: Attribution of blame – Vicarious victim *n* = 66, 10.4%	OHS containing a negative stereotype of individuals (a) belonging to the same age cohort and (b) with the same gender identity, as the participant. Attribution of blame claim present, blaming the group that is targeted (it is their own fault that they are targeted).	2a + *I'm posting this comment because young people have themselves to blame for being ridiculed online.*	1b + *I'm posting this comment because females have themselves to blame for being ridiculed online.* Or (depending on the participants own gender identity): 1b + *I'm posting this comment because males have themselves to blame for being ridiculed online.*

Note. OMD stands for Online Moral Disengagement. Participants were assigned randomly to one of the ten conditions and are shown both vignettes. After each vignette, participants were asked to label the vignette (open question) and to rate items on the perceived offensiveness, harmfulness to society, need to intervene, and intention to intervene. Participants who identify as 'non-binary' (n = 9) were randomly assigned to one of the ten conditions, similar to female and males. However, their responses on vignette B were not included in the analyses (as they do not identify themselves as male or female).

Appendix B

Table B1. *Categories of Labels That Resulted After Open-Coding and Refining.*

Category	Definition used	Examples of answers (Dutch)	Examples of answers (translated from Dutch to English)	% (*N*) of participants that provided this label to …	
				OHS targeted on age	OHS targeted on gender identity
1.OHS	The expression towards an individual or group of “hatred or degrading attitudes toward a collective” (Hawdon et al., 2017, p. 254), aiming to devalue and demean them collectively	haatspraak, haatbericht , hatend, vijandig, seksistisch, stereotiep, generaliserend, discriminerend, misogynistic, racistisch, over een kam scherend, aanval op een groep, mensen in een hokje plaatsen, vooroordelen, onrechtvaardig, discriminerend	hate speech, hate message , hateful, hostile, sexist, stereotypical, generalizing, discriminatory, misogynistic, racist, lumping, attacking a group, pigeonholing people, prejudice, unjust, discriminatory	16.4 (44)	57.2 (199)
2. Online aggression	Intentionally delivering harm through ICT to a person or a group of persons who perceive(s) such acts as offensive, derogatory, harmful, or unwanted (Grigg, 2010, p. 152).	agressief, aanvallend, kwetsend, dreigend	aggressive, offensive, abusive, threatening	2.2 (6)	5.5 (19)
3. Online incivility	Speech that violates politeness norms by using rude, vulgar, and/or disrespectful language (Kümpel & Unkel, 2023).	onbeleefd, onprettig, onaardig, beledigend, spottend, brutaal, heftig, ongepast/niet passend, opstandig, grof, asociaal, onredelijk, gemeen, onvriendelijk, onbeschoft, ondoordacht, kortzichtig, schaamteloos, niet normaal, asociaal, hard, schokkend, bot, onfatsoenlijk, neerbuigend, schrikwekkend, schokkend, onaanvaardbaar, minachtend, respectloos, aanstootgevend, hard	rude, unpleasant, unkind, insulting, mocking, insolent, vehement, inappropriate/not appropriate, rebellious, rude, antisocial, unreasonable, mean, unfriendly, rude, inconsiderate, short-sighted, shameless, not normal, antisocial, harsh, shocking, blunt, indecent, condescending, frightening, shocking, unacceptable, disdainful, disrespectful, offensive, harsh	40.9 (110)	21.6 (75)
4. Negative online reaction	Speech that is perceived as negative, without mentioning elements that are included in the definitions of OHS, online aggression and online incivility.	negatief, niet fraai, kinderachtig, naar, vervelend, onnodig/niet nodig, niet leuk, minder prettig, niet mooi	negative, not nice, childish, nasty, annoying, unnecessary/not necessary, not fine, less pleasant, not pretty	18.6 (50)	9.8 (34)
5. Freedom of speech/ neutral reaction	Freedom of speech: Speech that is perceived as the expression of one’s opinions and beliefs in a way that is allowed/will not lead to punishments by the government. Neutral: Speech that does not express an opinion or belief.	vrijheid van meningsuiting, eigen mening, aanvaardbaar, normaal, oké, neutraal, prima	freedom of expression, own opinion, acceptable, normal, okay, neutral, fine	16.4 (44)	3.7 (13)
6. Positive online reaction	Speech that is perceived as positive, because, for instance it is perceived as entertaining or prosocial.	positief, grap, goed, leuk, aardig, ludiek	positive, joke, good, fun, nice, playful	5.6 (15)	2.3 (8)
N (total)				269	348

Appendix C

Full pretest results

In order to select reliable stimulus materials, a pretest survey was administered among 18- to 25-year-olds, recruited via the institution's human subject pool (university students). In total, the pretest sample consisted of *N* = 134 with a mean age of 21.11 (SD_{age} = 1.99) and 73.1% (*N* = 98) identified as female, 26.1 % (*n* = 35) as male, and 0.8% (*n* = 1) with another gender identity. The majority of the pretest sample had the Dutch nationality (82.84%, *n* = 111), other nationalities included were Polish (3.73%, *n* = 5), Turkish (2.24%, *n* = 3), Romanian (1.49%, *n* = 2), Chinese (1.49%, *n* = 2) and others (8.33%, *n* = 11). Materials were tested for personal and societal relevance, perceived realism, connectedness with the target group, perceived presence of a negative stereotype, and the ideological leaning of the comment.

Negative stereotypes of individuals with a certain age and gender identity were pretested. For each group characteristic, a negative stereotype of the ingroup (young people; males/females) and outgroup (older people; males/females) were tested for two different discussion topics (the target group's driving skills and leadership qualities). Pretest respondents were exposed, at random, to one negative stereotype (either about the ingroup age, ingroup gender, outgroup age, or outgroup gender) per discussion topic. The discussion topics were assessed based on personal relevance and societal relevance (based on Skitka et al., 2017), both measured on a scale from (1) *strongly disagree* to (7) *strongly agree*. The negative stereotypes on age and gender were assessed on perceived realism, target group attachment, perceived negative stereotype, and the ideological leaning of the comment. Perceived realism was measured with two items: *Do you think people on social media discuss this issue like displayed here or in a similar manner?*, (1) *not at all like this* to (7) *very much like this*; *How likely is that you would come across a discussion like this on social media?'* (1) *very unlikely* to (7) *very likely*. Target group attachment was measured with one item: *I feel strongly connected to the group that is being targeted in the comment*; (1) *strongly disagree* to (7) *strongly agree*. Perceived negative stereotype was also measured with one item: 'The comment contains a negative stereotype about the group that is being targeted'; 1 (strongly disagree) to 7 (strongly agree). Finally, pretest respondents also indicated the ideological leaning of the comment (1 left-leaning to 10 right-leaning).

Both discussion topics (driving skills and leadership qualities) were perceived as personally and societally relevant and were, therefore, perceived as suitable for the present study. A paired samples *t*-test *t*(133) = −0.24, *p* = .811), showed no significant differences for personal relevance of the discussion topics. However, for societal relevance, a significant difference *t*(133) = −6.15, *p* < .001), was found with leadership qualities perceived as more relevant for society (*M* = 5.63, *SD* = 1.21), compared to driving skills (*M* = 4.81, *SD* = 1.20).

The negative stereotypes on age and gender were assessed on perceived realism, target group attachment, perceived negative stereotype, and the ideological leaning of the comment by comparing mean scores on these items for the four groups (ingroup age, ingroup gender, outgroup age, and outgroup gender) by means of two one-way ANOVA tests (one for each discussion topic). For the discussion topic driving skills, no differences were found between groups for realism and all negative stereotypes were perceived as (somewhat) realistic (range *M* on these two items: 4.50 – 5.70. For target group attachment, significant differences were found *F*(3,130) = 7.846, *p* < .001, but only for ingroup versus outgroup (for both gender identity and age), but participants felt equally connected to individuals with the same age and individuals with the same gender identity. No significant differences were found for the perceived presence of a negative stereotype, respondents (somewhat) agreed with the presence of a negative stereotype (range *M* = 5.66–6.27). Finally, no significant differences were found for the ideological leaning of the stereotypes. As the present study aimed to select a targeted group for which the perception of the OHS against this group is not determined by one's political orientation, the average ideological leaning of the stereotypes should be ideally as close to the midpoint of the scale as possible (5). This was true for all four groups (range *M* = 5.42–5.90). Based on these results, both gender and age seem suitable as target group characteristic for the vignette with discussion topic 'driving skills'.

For the discussion topic leadership qualities, for the first item regarding perceived realism (*Do you think people on social media discuss this issue like displayed here or in a similar manner?*) no differences were found between the groups and for all groups the negative stereotype was perceived as somewhat realistic (range *M* = 4.37–5.16).

For the second item regarding perceived realism (*How likely is that you would come across a discussion like this on social media?*), significant differences were found *F*(3,130) = 4.019, *p* = .009; the outgroup gender negative stereotype was perceived as less realistic (*M* = 4.24, *SD* = 1.85) compared to the negative stereotypes of the other groups, that were perceived as, on average, somewhat realistic (range *M* = 5.11–5.38). For target group attachment, significant differences *F*(3,130) = 22.428, *p* < .001 were only found for ingroup versus outgroup (for both gender identity and age), but participants felt equally connected to individuals with the same age and individuals with the same gender identity. No differences were found for the perceived presence of a negative stereotype: for all targeted groups, respondents (somewhat) agreed with the presence of a negative stereotype (range *M* = 5.65 – 6.28). Finally, significant differences were found for the ideological leaning of the stereotypes *F*(3,130) = 3.449, *p* = .019. As the present study aimed to select a targeted group for which the perception of the OHS against this group is not determined by one's political orientation, the average ideological leaning of the stereotypes should be ideally as close to the midpoint of the scale as possible (5). This was true for both the ingroup age (*M* = 5.70, *SD* = 2.01) and outgroup age stereotype (*M* = 5.41, *SD* = 1.65), but not for the ingroup gender stereotype (more right-leaning: *M* = 6.13, *SD* = 3.01) and for the outgroup gender stereotype (more left-leaning: *M* = 4.46, *SD* = 2.33). Based on these results, age was selected as target group characteristic for the vignettes with discussion topic 'leadership qualities' and gender identity was selected as target group characteristic for the vignettes with discussion topic 'driving skills'.

A final goal was to pretest online moral disengagement strategies statements. For each strategy (moral justification, diffusion of responsibility, distortion of consequences, and attribution of blame) two statements were tested, in order to select one statement for each strategy to include in the experiment. Participants were asked how likely they think it is that people post negative reactions about others online for each statement representing an online moral disengagement strategy. The OMD statements were constructed based on the Moral Disengagement through Technology Questionnaire of Marín-López et al. (2020). Respondents indicated on a 7 point-Likert scale: (1) *very unlikely* to (7) *very likely*, how likely they thought that people post negative reactions about others online for each of the two statements representing each of the four online moral disengagement strategies (eight statements in total). For each strategy, the statement with the highest mean score (most likely) was selected for the experiment, as these statements were, according to the pre-test participants, the most plausible to encounter online. All selected statements had an average score *M* ≥ 5.40, meaning they were perceived as (somewhat) likely. The selected statements also differed, based on paired samples *t*-tests, significantly from the other statement that represented the same strategy.

About Author

Sara Pabian is an Assistant Professor at the Department of Communication and Cognition of Tilburg University and is also an affiliated researcher of the Department of Communication Studies of University of Antwerp. Her research focuses on risks and opportunities of online interactions among both adolescents and adults.

https://orcid.org/0000-0001-9676-7553

✉ **Correspondence to**
Sara Pabian, Tilburg University, Tilburg center for Cognition and Communication, Warandelaan 2, 5037 AB Tilburg, Netherlands, University of Antwerp, Department of Communication Studies, Sint-Jacobstraat 2, 2000 Antwerp, Belgium, s.j.r.pabian@tilburguniversity.edu