Unveiling the Determinants of Bystander Intervention: Exploring the Impact of Victim Anonymity and Social Group Membership on Bystanders' Intention to Help Victims of Online Hate Speech

Daphne Zilverberg
SNR 2060695

Master's Thesis
Communication and Information Sciences
Specialization Business Communication and Digital Media

Department Communication and Cognition
School of Humanities and Digital Sciences
Tilburg University, Tilburg

Supervisor: Dr. A. Schouten
Second Reader: Dr. S. Pabian

June 2024

Technology statement

The technologies used in the writing process were the Academic Phrasebank and Grammarly to make suggestions for paraphrasing. In addition, Grammarly was used to check the English grammar.

Abstract

Online hate speech is a prevalent issue, inflicting negative effects on the emotional and physical well-being of victims. Because the support of bystanders can stop the bullying, bystanders play a pivotal role in protecting the well-being of victims. This study aimed to examine how the victim's level of anonymity (anonymous vs. identifiable) on social media affects bystanders' attitudes and intention to intervene, using passive or active interventions. Additionally, this study examined whether bystanders' attitudes or intentions to help the anonymous or identifiable victim varied based on the social group (in-group vs. out-group) to which the victim belonged. It was expected that bystanders have a more favorable attitude and an increased intention to help a victim when the victim is identifiable, through increased feelings of social presence and social attraction. Additionally, it was expected that bystanders have a more favorable attitude and a stronger intention to help a victim when the victim is anonymous and part of the in-group, compared to when the victim is identifiable and part of the in-group. The results of the experiment, with 160 participants, showed that bystanders have a more favorable attitude towards helping a victim when the victim is identifiable rather than anonymous due to increased feelings of social attraction. However, since social attraction and social presence were not found to mediate the relationship between the level of anonymity of the victim and the actual helping intentions of bystanders, more research is needed to examine the relationship between anonymous vs. identifiable victims and bystander intervention. The findings of this study may help victims of online hate speech on how to present themselves on social media to garner bystander support.

Keywords: online hate speech, anonymity, social group membership, social attraction, social presence, group identification, bystander intervention

Table of contents

Introduction

Surveys on the prevalence of online hate speech across various countries show that online hate speech is present everywhere: 42%-67% of young adults (18 to 25 years old) from various countries have witnessed hate speech online (Keipi et al., 2017) and 21% have even experienced being a victim of online hate speech themselves (Räsänen et al., 2016). Online hate speech has a negative impact on people's emotional state, as it can lead to symptoms of anxiety, depression (Wachs et al., 2022), and even suicidal thoughts (Chaudhary et al., 2021). In addition, online hate speech can also limit the diversity of opinions online, as it discourages vulnerable groups to engage in online discussions and express their beliefs due to fear of harassment (Munger, 2016). Although online hate speech is defined as attacking a group or an individual online because of, for example, their gender, race, or sexual orientation, this study focuses on hate speech targeted towards a specific individual (Castaño-Pulgarín et al., 2021).

Unlike offline hate speech, online hate speech can cause more damage to the victim because incidents can be witnessed by a larger audience (Schacter et al., 2016). As a result, the responses of bystanders play a crucial role in shaping the victim's well-being. In the context of online hate speech, bystanders can be defined as the individuals who observe the bullying situation (Zeng et al., 2023). Positive responses of bystanders, such as defending the victim, can lessen the victim's suffering, while negative responses, such as doing nothing, can empower the bully to continue and increase the victim's suffering (Patterson et al., 2016).

Various social media platforms provide users varying levels of anonymity, which can influence bystanders' intention to intervene (You & Lee, 2019). However, little is known about bystander behavior regarding the victim's anonymity. Anonymity refers to the ability to remain unidentified to others (Brody & Vangelisti, 2015). Bystanders may find it hard to empathize with anonymous victims, as it is difficult to feel a connection with a victim when the victim does not share any information about his or her identity (Small & Loewenstein,

2003). Indeed, some research shows that being anonymous as a victim can result in bystanders not helping the victim, as the anonymity reduces the feeling of responsibility (Yam & Reynolds, 2014). Conversely, identifiable victims do display personal information on their social media accounts, such as a profile photo, a name, or their gender (Kogut & Ritov, 2005). Due to these human-like cues, bystanders may find it easier to view the victim as a real person with emotions, who is worthy of receiving help (Lee & Feeley, 2016).

Studies on bystander behavior also show the importance of shared group membership (Forbes et al., 2019; Levine et al., 2002). More specifically, research suggests that when individuals belong to the same social group as the victim, it tends to have a positive effect on their attitudes and behavior towards helping the victim. According to the social identity theory by Tajfel and Turner (1979, as cited in Hornsey, 2008), people tend to categorize themselves and others in groups based on characteristics they share. These groups are referred to as in-groups and out-groups. In-groups consist of individuals who share a common identity, such as a university affiliation or religion. Out-groups consist of individuals that a person lacks identification with and does not share similar views with. As a result, people tend to dislike individuals who do not belong to their group. Previous research has established that when bystanders recognized a victim who was physically abused as a fellow classmate, so a member of the in-group, bystanders were more inclined to support the victim (Levine et al., 2002).

Although existing research has individually examined the influence of anonymity and group membership on bystander behavior, these factors have not yet been combined with each other within the research field of online hate speech. While research suggests that bystanders find it easier to empathize with identifiable victims than anonymous victims (Small & Loewenstein, 2003), it is expected that when the anonymous victim is perceived to be part of the same social group as the bystander, bystanders' intention to help the victim will be

stronger than when the victim is identifiable and part of the in-group. This expectation is based on the findings of the Social Identification Model of Deindividuation Effects (SIDE), which argues that anonymity conceals personal characteristics causing bystanders to perceive fewer differences between themselves and the victim (Lea & Spears, 1992). As a result, bystanders may experience a stronger sense of similarity with the victim than when the victim is identifiable. This perceived similarity plays an important role in the behavior of bystanders (Postmes et al., 2001).

Understanding the relationship between the factors level of anonymity and social group membership can provide new insights into why some bystanders intervene in situations of online hate speech while others do not. This study aims to investigate how these factors influence bystanders' attitudes towards helping the victim and bystanders' intention to intervene in situations of online hate speech, to contribute to research aimed at minimizing or preventing the occurrence of hate speech online. Therefore, the following research question is formulated:

> RQ: How does the level of victim anonymity (anonymous vs. identifiable) influence the attitude and intention of bystanders between the ages of 18 and 25 years old to help a victim in situations of hate speech online, and how does this attitude and intention vary based on bystanders' perception of the social group to which the victim belongs?

Theoretical framework

Online hate speech

The purpose of online hate speech is to intentionally cause harm to the victim (Sellars, 2016). Online hate speech is usually communicated via social media, environments such as online gaming, or other mediated channels of communication (Corazza et al., 2020), and can be communicated using, for example, memes, videos, or text (Schmid et al., 2022). Because the platforms through which online hate speech is communicated are usually public spaces, such as social media, online hate speech is public in nature (Sellars, 2016). As a result, one single incident of online hate speech can cause significant and immediate harm to the victim, as it can lead to repeated victimization of victims due to the large public reach on social media. This differentiates online hate speech from other types of insulting communication such as cyberbullying, as cyberbullying involves long-term exposure to a recurring pattern of bullying behavior (Tokunaga, 2010).

Another characteristic that differentiates online hate speech from cyberbullying is that online hate speech is targeted at a group or an individual who is perceived as someone who represents that group (Sellars, 2016). Therefore, unlike cyberbullying, online hate speech is characterized by the purpose of humiliating individuals based on their membership to a certain social, religious, or racial group (Keipi et al., 2017). Thus, online hate speech targets victims based on the social identity of the victim (Major & O'Brien, 2005). As a consequence, research finds that being a victim of online hate speech cannot only lead victims to question their own self-worth but also their value of belonging to a certain social group (Costello et al., 2019).

Social media is considered to encourage the spread of hate speech for numerous reasons. Firstly, the immediate and rapid nature of social media causes users to impulsively comment hate (Brown, 2017). Second, the social media features that allow users to share

content with just one click, such as ‘retweet or ‘share’, facilitate the fast spread of online hate speech to a large audience of people, which can lead to repeated victimization (Benigni et al., 2017). Third, in an online environment, negative information disseminates more rapidly than positive information. The findings of Tsugawa and Ohsaki (2015) revealed that negative information is 1.2-1.6 times more reposted than positive information and that negative information is 1.25 times quicker disseminated. Lastly, social media offers users the ability to be anonymous, which can cause users to say things they would not say in real life because they feel like they can express themselves more freely (Suler, 2004).

The role of bystanders in online hate speech

Research on hate speech reports that bystanders play a pivotal role in the bullying process (Obermaier et al., 2023). This is because bystanders can have a significant influence on the severity of the bullying and the effect of the bullying on the victim by taking on various roles (Tsang et al., 2011). These various roles are: supporting the bully, doing nothing, or helping the victim (Salmivalli et al., 1996). When bystanders support the bully, by helping the bully or showing the bully that they accept the behavior, they praise the bully for his or her actions, which influences the bully to become more violent and perform the behavior more frequently (Salmivalli, 2010). In addition, doing nothing can also lead the bully to believe that the bystander accepts the behavior (Salmivalli et al., 1996). In contrast, bystanders can also choose to engage in positive bystander behavior by choosing to defend the victim. When bystanders decide to defend the victim, they jeopardize the status of the bully and, as a result, the bully may make the decision to stop posting hateful comments (Hawkins et al., 2001).

Positive bystander intervention can be classified into two groups, namely active or passive intervention (Obermaier, 2022). Active intervention refers to bystanders taking the responsibility to support the victim upon themselves. Examples of active intervention on

social media are confronting the bully or supporting the victim through private messages (Brody & Vangelisti, 2015). On the other hand, passive intervention refers to bystanders reaching out to other people to deal with the situation. Examples of passive intervention are flagging a post as inappropriate to bring it to the moderator's attention or reporting accounts (Crawford & Gillespie, 2014). Research shows that active interventions, specifically messages from bystanders including moral support, are the most helpful for victims (High & Young, 2018). However, less substantial actions such as liking a post of the victim are also considered to be an effective intervention, because by liking a victim's post bystanders can signal moral support to the victim (Ellison et al., 2014).

In order to explain when bystanders decide to intervene in situations of online hate speech, the bystander intervention model of Latané and Darley (1970) can be used. According to this model, bystanders are more likely to intervene in situations of hate speech when they recognize a situation as critical, perceive the situation as dangerous and threatening for the victim, believe that they are responsible for helping the victim, and eventually decide to support the victim (Obermaier et al., 2023). In the literature, it has been found that bystanders believe that they are responsible to help the victim and, as a result, reported a higher willingness to help the victim when there is a limited amount of other bystanders nearby who are able to help the victim (Obermaier et al., 2016).

In addition, researchers have looked into several other individual, situational, and contextual factors that increase bystanders' intention to intervene in situations of online hate speech. For example, a lower risk of danger for the bystander can motivate bystanders to help victims (Rovira et al., 2021). Furthermore, on an individual level, confidence in one's abilities (DeSmet et al., 2014), empathy (Abbott & Cameron, 2014), and self-control (Erreygers et al., 2016) have been found to increase bystander intervention. Additionally, Kunst et al. (2021) showed that solidarity norms, an individual's belief that people who are good care for the

well-being of other people, encourage positive bystander behavior. Lastly, experiences with prior victimization (Cascardi et al., 2018) and low moral disengagement attitudes, such as not blaming or dehumanizing the victim (DeSmet et al., 2014), have been found to positively influence bystanders' intention to intervene.

Anonymous vs. identifiable victims on social media

Another factor that can influence bystanders' willingness to intervene is the degree to which the victim is visible on social media (Schacter et al., 2016). Social media offers users the ability to experiment with different levels of visibility (Aizenkot, 2020). Users can choose to share their identity through their profile or they can choose to remain anonymous. Anonymity relates to the extent to which victims can conceal their identity from others (Brody & Vangelisti, 2015). According to Kogut and Ritov (2005), the anonymity of the victim makes it hard for bystanders to visualize the victim experiencing pain. This can lead to decreased feelings of compassion for the victim, as bystanders cannot engage in perspective-taking behavior. In addition, the anonymity of the victim can cause bystanders to perceive the victim as less human, which, as a result, may lead bystanders to perceive the victim as less affected by the hate speech (Yam & Reynolds, 2014). Therefore, the choice to remain anonymous as a victim can lead bystanders to take incidents of hate speech less seriously.

These findings align with the identifiable victim effect, which suggests that people are willing to offer more support to identifiable victims compared to unidentifiable victims (Jenni & Loewenstein, 1997). Jenni and Loewenstein (1997) explained that identifiable victims are more vivid than unidentifiable victims. Vividness entails that details about the victim in the newspaper or on television, such as pictures, a name, or other personal information, are provided. Seeing a person's name, picture or other personal information can humanize them which can induce empathetic feelings (Xu & Lombard, 2017). Additionally, these vivid details can cause people to feel a sense of familiarity with the victim, which may increase the

perceived importance of taking action to help the victim. As stated by Schelling (1968, as cited in Jenni & Loewenstein, 1997), the more people know about a person, the more people care about the person. So, specific and graphic information about the identity of a victim rather than unidentifiable information has more influence on people's behavior (Zhao et al., 2024). Therefore, in the context of online hate speech, bystanders may be more likely to help victims when they are identifiable rather than when they are anonymous.

Two processes may explain why people are less likely to help an anonymous victim. First, bystanders may feel less socially attracted to anonymous victims than identifiable victims. Social attraction refers to the degree to which an individual likes another individual and perceives the other individual as a potential friend (McCroskey & McCain, 1974). Previous research has found that when a victim can be personally identified, rather than when a victim lacks identifiable features, bystanders are more likely to have a positive feeling of liking towards the victim (Sassenberg & Postmes, 2002). This attraction, in turn, can lead to bystanders feeling more compassion for the victim (Hill & Courtright, 1981) and even an increased motivation to provide support (Graf & Riddell, 1972). According to research, the attraction towards another individual can be caused by feelings of familiarity (Moreland & Zajonc, 1982). As previously stated, it is easier for bystanders to perceive a sense of familiarity with identifiable victims as the identifiable cues, such as a human-like profile picture, can remind the bystanders of people that they know (Jenni & Loewenstein, 1997). Therefore, it is expected that bystanders are more inclined to like identifiable victims, compared to anonymous victims.

Second, social presence may explain why bystanders are less likely to help anonymous victims than identifiable victims. Social presence is defined by Short et al. (1976) as the degree to which people are aware of the presence of the other person in an online environment, and the degree to which the other person is recognized as a real person. As

identifiable victims share personal information on their social media profile, they are more likely to be perceived as a real person (Jones, 1991). In addition, the ability to see a human-like profile photo may reduce the feeling of psychological distance between the bystander and the victim, as being able to create a visual perception of the victim can increase the perception of closeness (Rim et al., 2014). Indeed, several studies show that people who use a human-like profile photo on their social media account, increase the feeling of social presence in an online environment (Newberry, 2001; Xu, 2014). This can result in bystanders feeling more emotionally connected to the victim, which makes them more likely to help the victim (Xu & Lombard, 2017). So, using identifiable human-like cues can lead bystanders to feel psychologically closer to the victim and perceive the victim as more 'real' (Lesner & Rasmussen, 2014), which can increase bystanders' feelings of responsibility to intervene (Jones, 1991).

Based on these findings, it is expected that when a victim of online hate speech is anonymous compared to identifiable, bystanders may feel less attracted to the victim and perceive the victim as less socially present, resulting in bystanders being more likely to ignore the bullying situation. Therefore, the following hypothesis is formulated:

> H1: When the victim is anonymous, bystanders feel (a) less attracted to the victim and (b) less social presence, resulting in a less favorable attitude of bystanders towards helping the victim and a reduced intention to help the victim, compared to when the victim is identifiable.

Social group membership and bystander intervention

As a species, humans have a natural tendency to form and identify with various groups (Rusch, 2014). The social identity theory by Tajfel and Turner (1979, as cited in Everett et al., 2015) suggests that when individuals perceive themselves to be part of a certain group, they

are likely to identify themselves in a way that aligns with the identity of the group. As a result, individuals experience a sense of a shared identity with other members of the group (Ellemers et al., 2002). Social groups can be formed based on various criteria, such as nationality, ethnicity, or religion (Everett et al., 2015). Following the process of individuals categorizing themselves into groups and identifying with this group, they start to compare their own group to other groups (Hornsey, 2008). The group that consists of members with whom an individual shares similar interests, is called the in-group. On the other hand, the out-group consists of people an individual does not identify and share similar interests with. So, people tend to view their own group positively by making comparisons with other groups and view them negatively. As a result, people are likely to treat in-group members in a more favorable way than out-group members (Everett et al., 2015).

Research on bystander behavior shows that shared group membership between a bystander and a victim can influence bystanders' behavior (DeSmet et al., 2014). For example, research by Levine et al. (2005) found that bystanders are more likely to support an injured victim who is perceived to be a fan of the same football team (in-group) rather than a different football team (out-group). So, categorizing a victim as a member of the in-group and perceiving the victim to be endangered induces prosocial behavior (Dovidio et al., 1997; Penner et al., 2005). More specifically, the perception of belonging to the same social group which fosters the feeling of a shared identity, can result in increased feelings of empathy and concern for the welfare of the members of that group (Miyazono & Inarimori, 2021). Moreover, research shows that higher levels compared to lower levels of identification with a group increases the feeling of connection with the members of that group (Doosje et al., 1995) and, as a result, increases the likelihood of supporting these members (Ellemers et al., 1997). Therefore, the following hypothesis is proposed:

H2: Bystanders are more likely to identify with a victim of online hate speech who belongs to the in-group compared to the out-group and, as a result, have a greater intention to help the victim of the in-group and a more favorable attitude towards helping the victim.

Anonymous vs. identifiable victims and social group membership

Although bystanders might be more likely to help victims who are identifiable rather than anonymous, it is unclear whether this intention to help the victim changes when the bystanders perceive the victim as belonging to the in-group compared to the out-group. While this study expects that bystanders are more likely to help identifiable victims than anonymous victims, it is expected that when the anonymous victim is perceived to be part of the in-group the intention of bystanders to help the victim will be stronger than when the victim is identifiable and part of the in-group. This can be explained by the Social Identification Model of Deindividuation Effects (SIDE). The SIDE model explains how online users overestimate similarity in an anonymous context as anonymity conceals personal characteristics and thereby interpersonal differences (Lea & Spears, 1992). As a result, in these anonymous circumstances, users can feel more equal and connected to each other than would have been the case if the users were identified (Lea et al., 2001).

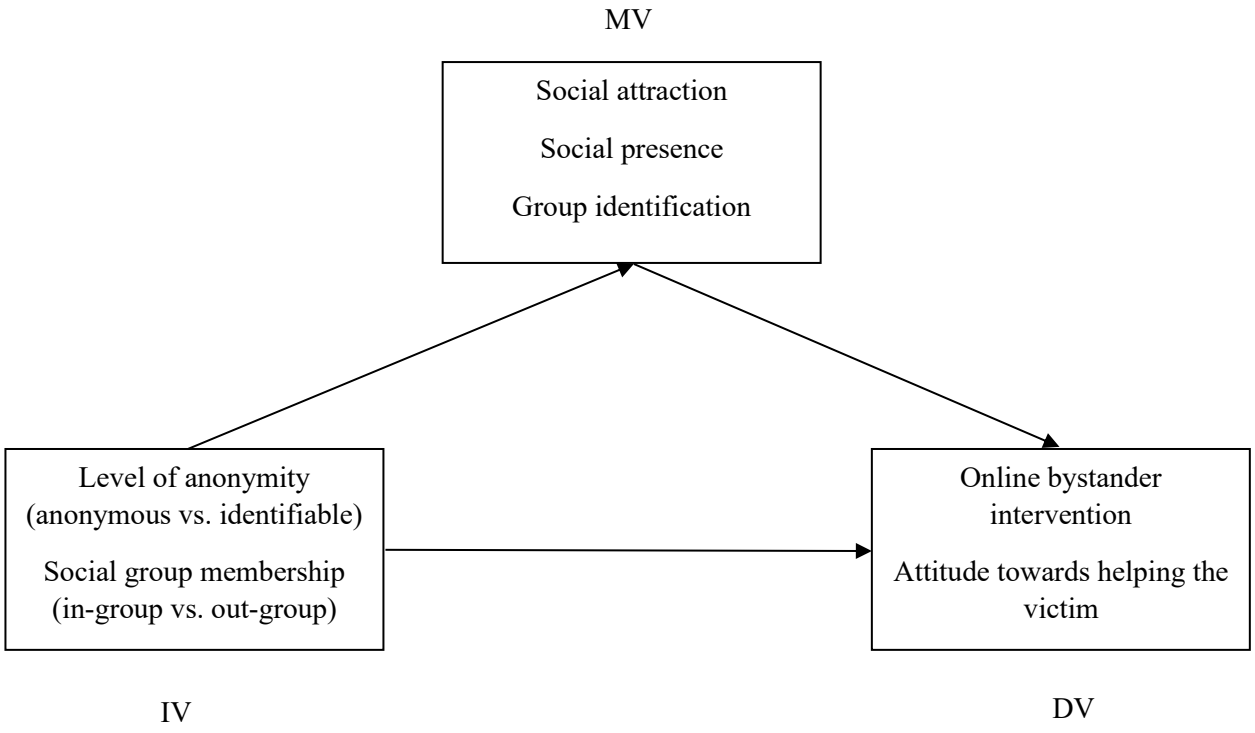
The process of a reduced attention to interpersonal differences and an increased attention towards a shared identity is called ‘depersonalization’ (Kim & Park, 2011). When bystanders have limited information about the anonymous victim, any cue present can become important in shaping a perception of the victim. For example, if solely information such as ethnicity is provided without any other personal information, bystanders might view the victim as a prototypical group member (Postmes et al., 1998). When the victim’s only available cue, such as ethnicity, resembles the ethnicity of the bystander, it implies they share

a social identity (Turner 1984, as cited in Kim & Park, 2011). This feeling of a shared identity increases a sense of connection (Kim & Park, 2011). On the contrary, when a victim is identifiable, more personal information is displayed, which increases the chance of bystanders perceiving differences between themselves and the victim. These perceived differences interfere with the process of identification (Lee & Nass, 2002). Therefore, following the SIDE model, the subsequent hypothesis is formulated:

> H3: Bystanders have a more favorable attitude towards helping a victim and are more likely to help a victim when the victim is anonymous and part of the in-group, compared to when the victim is identifiable and part of the in-group.

Figure 1

Conceptual model



Methodology

Participants

This study concerns a quantitative research in which participants completed an online questionnaire via the platform Qualtrics. Since, as indicated in the introduction, young adults often encounter hate speech online and even experience being a victim of online hate speech themselves (Keipi et al., 2017), participants had to be between the ages of 18 and 25 years old. The participants were recruited through convenience sampling using the social network of the researcher. All participants needed to have the Dutch nationality and had to be students from Tilburg University, as this was used to manipulate shared identity. In total, the questionnaire was completed by 222 participants. 62 participants were removed from the sample as they did not meet the age requirements, refused to give informed consent, did not study at Tilburg University, or did not finish the questionnaire. In the end, the data analysis included a final sample of 160 participants. The mean age of the participants was 22.6 years (*SD* = 1.37*).* Of all the participants, 71.9% were female and 28.1% were male. In addition, the highest level of education completed or the current education reported by the participants was 1.3% high school, 11.3% HBO bachelor's degree, 1.3% HBO master's degree, 25.6% WO bachelor's degree, 60% WO master's degree, and 0.6% selected the response option 'other'.

Design

To study the research question, an experimental, between-subject design was used. As shown in Table 1, this study included a 2 x 2 factorial design, with level of anonymity (anonymous vs. identifiable) and social group membership (in-group vs. out-group) as independent variables. The dependent variables were the behavioral intention of the bystanders to help the victim targeted by online hate (Obermaier, 2022) and the attitude of the bystanders towards helping the victim (Webb et al., 2000). The mediating variables in this

study were social presence (Weidlich et al., 2018), social attraction (McCroskey & McCain, 1974), and group identification (Doosje et al., 1995).

Table 1

2 x 2 Factorial design

		Level of anonymity	
		Anonymous	Identifiable
Social group membership	In-group	Online bystander intervention & attitude towards helping the victim	Online bystander intervention & attitude towards helping the victim
	Out-group	Online bystander intervention & attitude towards helping the victim	Online bystander intervention & attitude towards helping the victim

Manipulations

In order to evaluate whether the participants actually perceived the online hate speech comment as hate, a pretest was performed. Four different hate speech comments were created by using a hate speech dataset from Röttger et al. (2021) and were shown to four different individuals who did not participate in this study. The individuals were asked to rank the four various hate comments from least hateful to most hateful. Based on this, the most hateful comment was chosen for this study.

As hate speech is defined as attacking an individual or a group of individuals based on group characteristics such as nationality (Castaño-Pulgarín et al., 2021), the hate speech comment in this study was targeted towards Dutch people. The message of the hate speech comment was informal (e.g., hashtags, exclamation marks), to create a realistic rather than a

scripted hate comment. The experimental stimuli consisted of screenshots of Facebook posts that were created via the website zeoob.com (Zeoob, 2022) (see Appendix A). Participants in all conditions were exposed to the same hate speech comment (see Table 2).

Table 2

Hate speech incident on social media

Social media post victim
Had a great discovery at the market today! I was able to score a nice bag at a bargain price by negotiating cleverly. Who said negotiating was out of style?? #bargainhunting #smartshopping #winning
Online hate speech message
Typical Dutch People, always only concerned with money! Why am I not surprised that the Dutch are showing their greedy side again? I'd rather die than ever be friends with a Dutch person. It's really f*king pathetic to take money from people selling stuff on the market who already have so little to survive. You guys are f*king scum.

The level of anonymity of the victim on social media was manipulated by altering the profile photo and the name of the victim. In the anonymous condition, the social media profile of the victim was paired with a non-human-like profile photo, namely a default photo outlining the shape of a head in grey colors. In addition, in the anonymous condition, a non-human-like name was used (see Figure 2). In the identifiable condition, the social media profile of the victim was paired with a human-like name and a human-like profile photo (see Figure 3). The identifiable social media profile that was shown depicted either a male or female victim, depending on the gender reported by the participants in the survey. This was done to ensure that the group characteristic 'gender' did not interfere with the group characteristic 'university affiliation'.

Figure 2

Anonymous profile of the victim on Facebook

Note. The anonymous profile photo was collected from the website Adobe Stock.

Figure 3

Identifiable profile of the in-group victims on Facebook

Note. The identifiable profile photos were collected from the website Pexels.

Furthermore, to manipulate social group membership, the university affiliation of the victim was altered. University affiliation served as a suitable measure of social identification since all participants were Tilburg University students. Therefore, the experimental stimuli consisted of two distinct social groups based on university affiliation, namely Tilburg

University as the in-group, and University of Twente as the out-group. In the in-group condition, participants were exposed to a fictitious Facebook profile which displayed in the profile information that the victim went to school at Tilburg University. In addition, the Facebook profile contained cues to the university affiliation by changing the banner of the profile to an image of the Tilburg University campus. Conversely, in the out-group condition, the participants were exposed to a Facebook profile that contained personal information which displayed that the victim went to the University of Twente. In this condition, the banner of the profile was also altered, only now with an image of the campus of the University of Twente (see Appendix B).

Procedure

Before the participants could begin with the questionnaire on Qualtrics, the participants received information about the study and were asked to read the informed consent. After the participants read and agreed with the informed consent, demographic questions were asked. The demographic questions involved questions about age, gender, nationality, level of education, and which college institution they were currently registered in. If the participants were not students at Tilburg University, the participants could not partake in the study and were redirected to the end of the questionnaire. The participants that were not Dutch or not between the ages of 18 and 25 years old, were manually deleted from the sample after the data collection.

After the demographic questions, the participants were randomly assigned to one of the four conditions. At first, the participants were exposed to a Facebook profile, from either an anonymous or identifiable victim, who studies at Tilburg University or the University of Twente. Thereafter, the participants were exposed to a Facebook post of the victim which received a hate speech comment. After the participants were exposed to the Facebook profile and post, the participants were asked to indicate to what extent they perceived the victim to be

visible and to what extent they identified with the victim's social group (Tilburg University vs. Twente University), to check whether the anonymity and social group membership manipulations were successful. Subsequently, the participants were asked to assess to what degree they perceived the victim as 'real' and to what degree they 'liked' the victim by answering the statements of the social presence scale (Weidlich et al., 2018) and the social attraction scale (McCroskey & McCain, 1974). After the questions about social presence and social attraction, the participants were asked to indicate to what extent they identified with the victim displayed in the Facebook post. In the end, participants were asked to assess how likely they were to help the victim and to rate their attitude towards helping the victim.

Once the participants completed the survey, a debriefing was presented in which was explained that the study was about measuring whether bystanders of online hate speech are more likely to help a victim who is anonymous or identifiable, and if this intention to help varies based on whether the victim belongs to the same social group or a different social group as the bystander. Additionally, in the debriefing, the participants were reminded to not hesitate to contact the researcher for any further questions about the study. Altogether, the questionnaire took approximately five minutes to complete.

Operationalization

Independent variables

Social attraction. The social attraction between the bystander and the victim of online hate speech was measured by asking participants to answer three items. These items were obtained from the interpersonal attraction measurement by McCroskey and McCain (1974). The scale contained six items, however, three items were removed as these items were less applicable for this study. One of the items in the scale that was included in this study was reverse-coded: "He (she) just wouldn't fit into my circle of friends". However, for this research, the reverse-coded item was rephrased using positive wording to make the item

congruent with the other items on the scale. All the scales in this study are measured on a 7-point Likert scale ranging from *strongly disagree* (1) to *strongly agree* (7). The reliability of the scale was good, Cronbach's α = .87 (*M* = 3.96*, SD* = 1.19). The items of all the scales from this study can be found in Appendix C.

Social presence. Social presence was measured using six items based on a scale by Weidlich et al. (2018). This scale measures the extent to which participants perceive others in an online environment as 'real' and 'nearby'. The scale originally consisted of ten items, however, four items were not included in this study because when they were translated from English to Dutch, they resulted in similar translations as some of the included items. An example of an included item is: "I can form clear impressions of the victim". The reliability of the scale was good, Cronbach's α = .93 (*M* = 3.67*, SD* = 1.45).

Group identification. To measure the extent to which the participants identified with the victim of online hate speech, the group identification scale was used (Doosje et al., 1995). This scale consists of four items and an example of one of the items is: "I feel strong ties with the victim". One additional item was added to this scale to capture the group-level identification with the victim to a greater degree ("The victim and I belong to the same social group"). The scale demonstrated good reliability, Cronbach's α = .93 (*M* = 3.28*, SD* = 1.41).

Dependent variables

Online bystander intervention. The dependent variable 'online bystander intervention' was measured using a scale from Obermaier (2022). One of the items of the originally ten-item scale by Obermaier (2022) was excluded from this study because when translated to Dutch, it resulted in the same translation as one of the included items. Results from a factor analysis showed that the scale consisted of two separate dimensions, namely: passive intervention (items 1-3) (e.g., "I would flag the statements in a comment as 'hate speech") and active intervention (items 4-9) (e.g., "I would support the victim in a private

message"). The two factors explained 25.3% and 43.7% of variance. Although item 3 loaded slightly higher on the factor active intervention, this study decided to include it in the factor passive intervention as it made more conceptual sense. The factor loadings can be found in Appendix D. The separate dimensions that formed the bystander intervention scale had good reliability: passive intervention α = .75 (*M* = 4.21*, SD* = 1.51) and active intervention α = .89 (*M* = 2.55*, SD* = 1.20).

Attitude towards helping others. The dependent variable 'attitude towards helping others' was measured using a scale from Webb et al. (2000). The attitude of bystanders towards helping the victim was measured with three items (e.g., "For me, helping the victim is important"). The reliability of the scale was good, Cronbach's α = .74 (*M* = 4.72*, SD* = 1.08).

Manipulation check

Level of anonymity. A manipulation check was used to ensure that the participants accurately perceived the anonymous victim as anonymous and not identifiable. To check whether the manipulation of anonymity was successful, three items were presented to the participants (e.g., "The victim was identifiable to me"). The reliability of the manipulation check was good, Cronbach's α = .92 (*M* = 4.43*, SD* = 1.53).

Group identification. To check whether the participants identified with the in-group (Tilburg University) as opposed to the out-group (University of Twente), the participants were asked to rate how strongly they agreed with each item on the group identification scale of Brown et al. (1986). Five items of the originally ten-item scale were excluded as they were reverse-coded. The reverse-coded items were excluded to maintain consistency and to make it easier for participants to interpret the items. As a result, five items remained (e.g., "I am a person who identifies with Tilburg University"). Participants in the in-group condition were exposed to the group identification scale containing items related to identification with Tilburg University, and participants in the out-group condition were exposed to the group

identification scale containing items about identification with the University of Twente. Both scales had good reliability: identification with Tilburg University α = .91 (*M* = 5.27, *SD* = 1.13) and identification with the University of Twente α = .82 (*M* = 1.62, *SD* = 0.72).

Results

A Shapiro-Wilk test of normality was performed to discover whether the variables were normally distributed. The results showed that the distribution for all the variables departed from normality ($p < .05$). Based on this outcome, bootstrapping was used for the mediation analysis to acquire 95% confidence intervals. The interval was based on 1000 bootstrapped samples. While the individual variables were not normally distributed, the Shapiro-Wilk normality test of the MANOVA yielded a non-significant result ($W = .99$, $p = .078$), indicating that the assumption of normality was met for the MANOVA analysis.

Manipulation check

To discover whether the manipulations of anonymity and group identification were successful, an independent t-test was performed. The t-test revealed that participants perceived the anonymous victim ($M = 3.39$, $SD = 1.40$) as significantly different from the identifiable victim ($M = 5.39$, $SD = 0.90$), $t(127) = -10.7$, $p < .001$. Additionally, to test the manipulation of group identification the scores of the scale including items about the in-group and the scores of the scale including items about the out-group were combined into one variable. The t-test revealed that participants reported higher levels of identification with Tilburg University ($M = 5.40$, $SD = 1.13$) than the University of Twente ($M = 1.40$, $SD = 0.72$), $t(131) = -24.3$, $p < .001$. Thus, the manipulation of both anonymity and group identification was successful.

Test of hypotheses

In order to test the first hypothesis and the second hypothesis, a mediation analysis was performed. The mediation analysis was conducted in Jamovi using the MedMod module. Additionally, a Multivariate Analysis of Variance (MANOVA) was conducted to examine the main effects of the independent variables on the dependent variables and the interaction effects. Level of anonymity (anonymous vs. identifiable) and social group membership (in-

group vs. out-group) were inserted as the independent variables (IV), and bystanders' intention to intervene using passive or active interventions and bystanders' attitude towards helping the victim were inserted as the dependent variables (DV). Furthermore, social attraction, social presence, and group identification were included as the mediating variables (MV) in the MedMod model.

Hypothesis 1

The first hypothesis predicted that when a victim of online hate speech is anonymous, bystanders feel (a) less attracted to the victim and (b) less social presence, resulting in a less favorable attitude towards helping the victim and a reduced likelihood of bystanders' intention to help the victim, compared to when the victim is identifiable. The results of the MANOVA revealed that there was no main effect of level of anonymity on both the passive (F(1,156) = 0.67, p = .413, η_p^2 = .004) and active interventions to help the victim (F(1,156) = 0.00, p = .976, η_p^2 = .000), as well as the attitude towards helping the victim (F(1,156) = 0.03, p = .866, η_p^2 = .000).

When analyzing the indirect effects, the results of the mediation analysis revealed that social attraction does significantly mediate the relationship between the level of anonymity of the victim and the attitude towards helping the victim (β = .13, 95% CI [0.09, 0.52], z = 2.82, p = .005). Level of anonymity positively affects social attraction (β = .43, 95% CI [0.69, 1.37], z = 6.03, p < .001), and social attraction, in turn, positively affects attitude towards helping the victim (β = .30, 95% CI [0.08, 0.46], z = 3.19, p = .001). This finding indicates that bystanders are more likely to feel attracted to identifiable victims than anonymous victims and, as a consequence, have a more favorable attitude towards helping the victim. However, the results revealed that social attraction did not significantly mediate the relationship between level of anonymity and both the passive (β = .04, 95% CI [-0.14, 0.44], z = 1.07, p = .286) and active interventions to support the victim (β = .05, 95% CI [-0.07, 0.34],

z = 1.13, p = .260).

Additionally, the results of the mediation analysis revealed that social presence does not significantly mediate the relationship between level of anonymity and both the passive (β = .08, 95% CI [-0.20, 0.66], z = 1.31, p = .190) and active interventions to help the victim (β = .08, 95% CI [-0.10, 0.50], z = 1.31, p = .190), as well as the attitude towards helping the victim (β = -.06, 95% CI [-0.40, 0.16], z = -1.00, p = .320). Furthermore, no significant effect was found of social presence on the dependent variables passive (β = .14, 95% CI [-0.12, 0.38], z = 1.33, p = .185) and active interventions to help the victim (β = .14, 95% CI [-0.06, 0.29], z = 1.32, p = .185), and the attitude towards helping the victim (β = -.10, 95% CI [-0.22, 0.10], z = -1.00, p = .317). However, the results showed a significant effect of the independent variable level of anonymity on the mediating variable social presence (β = .58, 95% CI [1.28, 2.00], z = 8.92, p < .001). This finding indicates that the feeling of social presence increases when the victim is identifiable compared to anonymous.

Taken together, these findings show that solely social attraction acts as a mediator, and exclusively for the dependent variable attitude towards helping the victim. Therefore, the first hypothesis can only be partially supported. See Table 3 for the descriptive statistics of the mediating variables.

Hypothesis 2

The second hypothesis predicted that bystanders are more likely to identify with a victim of online hate speech who belongs to the in-group compared to the out-group and, as a result, have a greater intention to help the victim of the in-group and a more favorable attitude towards helping the victim. The results of the MANOVA revealed that there was no main effect of social group membership on both the passive (F(1,156) = 2.90, p = .090, η_p^2 = .018) and active interventions to help the victim (F(1,156) = 1.11, p = .295, η_p^2 = .007), and the attitude towards helping the victim (F(1,156) = 2.15, p = .144, η_p^2 = .014).

In addition, the mediation analysis revealed that group identification did not significantly mediate the relationship between social group membership and both the passive (β = .03, 95% CI [0.01, 0.31], z = 1.57, p = .116) and active interventions to help the victim (β = .05, 95% CI [0.00, 0.32], z = 1.73, p = .084), and the attitude towards helping the victim (β = .04, 95% CI [-0.00, 0.23], z = 1.66, p = .097). Additionally, the independent variable social group membership does not have a significant effect on the mediating variable group identification (β = .15, 95% CI [-0.02, 0.87], z = 1.87, p = .062). Therefore, it can be concluded that the second hypothesis is rejected. Nevertheless, the results did find a significant effect of the mediating variable group identification on the dependent variables attitude towards helping the victim (β = .28, 95% CI [0.10, 0.33], z = 3.63, p < .001) and the intention to help the victim through passive (β = .22, 95% CI [0.07, 0.42], z = 2.90, p = .004) and active interventions (β = .34, 95% CI [0.17, 0.43], z = 4.51, p < .001), indicating that the more bystanders identify with a victim, the more they are inclined to help this victim.

Table 3

Descriptive statistics of the mediating variables

	Social group membership	Level of anonymity	Mean	Std. Deviation	N
Social attraction	Out-group	Anonymous	3.31	1.153	38
		Identifiable	4.49	1.035	43
	In-group	Anonymous	3.55	0.938	39
		Identifiable	4.42	1.198	40
Social presence	Out-group	Anonymous	2.76	1.344	38
		Identifiable	4.55	1.012	43
	In-group	Anonymous	2.85	1.135	39
		Identifiable	4.38	1.273	40
Group identification	Out-group	Anonymous	2.39	1.249	38
		Identifiable	3.69	1.254	43
	In-group	Anonymous	3.26	1.218	39
		Identifiable	3.71	1.512	40

Hypothesis 3

To test the third hypothesis, a MANOVA was conducted. The third hypothesis predicted that bystanders have a more favorable attitude towards helping a victim and are more likely to help a victim when the victim is anonymous and part of the in-group, compared to when the victim is identifiable and part of the in-group. The descriptive statistics of all the conditions can be found in Table 4.

The univariate tests revealed that the interaction effect between the independent variables level of anonymity and social group membership was not significant for both the passive ($F(1,156) = 1.21$, $p = .273$, $\eta_p^2 = .008$) and active interventions of bystanders to help the victim ($F(1,156) = 1.56$, $p = .213$, $\eta_p^2 = .010$). Additionally, the interaction effect between the independent variables on the dependent variable attitude towards helping the victim was also found to be non-significant ($F(1,156) = 0.02$, $p = .895$, $\eta_p^2 = .000$). Therefore, it can be

concluded that social group membership does not influence the effect of anonymity or identifiability on helping intentions, indicating that the third hypothesis is rejected.

Table 4

Descriptive statistics of the MANOVA

	Social group membership	Level of anonymity	Mean	Std. Deviation	N
Indirect intervention	Out-group	Anonymous	3.96	1.466	38
		Identifiable	4.04	1.462	43
	In-group	Anonymous	4.64	1.454	39
		Identifiable	4.19	1.621	40
Direct intervention	Out-group	Anonymous	2.32	1.003	38
		Identifiable	2.56	1.060	43
	In-group	Anonymous	2.77	1.435	39
		Identifiable	2.53	1.249	40
Attitude towards helping the victim	Out-group	Anonymous	4.57	0.907	38
		Identifiable	4.63	0.921	43
	In-group	Anonymous	4.85	1.353	39
		Identifiable	4.86	1.119	40

Discussion

The goal of this study was to investigate whether the level of anonymity of the victim (anonymous vs. identifiable) influences the attitude and intention of bystanders between the ages of 18 and 25 years to help the victim of online hate speech. Additionally, this study investigated whether the attitude and intention of bystanders to help the anonymous or identifiable victim varied based on bystanders' perception of the social group (in-group vs. out-group) to which the victim belonged. Using an experiment, three hypotheses were tested. In the next paragraphs, the three hypotheses will be discussed in light of the research findings.

The first hypothesis predicted that when a victim is anonymous, bystanders feel (a) less attracted to the victim and (b) less social presence, resulting in a less favorable attitude towards helping the victim and a reduced intention to help the victim, compared to when the victim is identifiable. We found that social attraction, but not social presence mediated the relationship between the level of anonymity of the victim and bystanders' attitude towards helping the victim. This finding is consistent with that of Sassenberg and Postmes (2002) who found that people are more likely to feel attracted to a person who can be personally identified. A possible explanation behind this is that the use of identifiable cues, such as using a human-like profile photo and name on social media, can remind the bystanders of people that they know (Jenni & Loewenstein, 1997). This perceived familiarity, in consequence, can cause bystanders to have a more favorable attitude towards supporting the victim.

While the results showed a significant effect on bystanders' attitude towards helping the victim, no significant effect was found on their intention to help the victim. The results indicate that when a victim is identifiable, bystanders are more inclined to like the victim. Subsequently, when bystanders like a victim, bystanders are more likely to have a positive attitude towards helping that victim. However, liking a victim is seemingly not enough to provide actual support. This can be explained by the possibility that attitudes are more

susceptible to changes in the level of victim anonymity through the mediator social attraction than behavioral intentions. Therefore, although bystanders' attitudes may be positive, it does not automatically lead to increased intentions to help, as behavioral intentions are complex and influenced by various factors (Ajzen, 1991).

Furthermore, the results of the first hypothesis did not show a significant mediation effect of social presence on the relationship between the level of anonymity of the victim and bystanders' attitude and intention towards helping the victim. Since the results show that social presence and social attraction do not mediate the relationship between the level of anonymity of the victim and the actual intention of bystanders to provide support, other factors might be at play. For example, Zhao et al. (2024) found that people are more likely to support identifiable victims due to increased feelings of empathy. Nevertheless, the results did reveal a significant relationship between the independent variable level of anonymity and the mediating variable social presence. This finding indicates that bystanders are more aware of the presence of the victim in an online environment and are more likely to perceive the victim as a real person when the victim is identifiable. These results reflect that of Newberry (2001) who also found that people who use identifiable cues on their social media profile such as a human-like profile photo or name, increase the perception of social presence.

Although the results revealed a significant relationship between the independent variable and the mediating variable, no significant relationship was found between the mediating variable social presence and the dependent variables. A possible explanation for this might be that according to Kreijns et al. (2020), social presence can be distinguished into two dimensions, namely awareness of others and the feeling of psychological closeness with others. However, in this study, social presence was measured as a unidimensional factor. It is possible that the dimension of psychological closeness might have a more pronounced effect on bystander intervention since research by Biocca et al. (2001) argues that the feeling of

psychological closeness is more important as it creates a stronger sense of social presence than the mere awareness of others. Therefore, measuring these dimensions separately might reveal significant results.

The second hypothesis predicted that bystanders are more likely to identify with a victim who belongs to the in-group compared to the out-group and, as a result, have a greater intention to help the victim of the in-group and a more favorable attitude towards helping the victim. However, the results only found a significant relationship between bystanders' identification with the victim (mediating variable) and bystanders' attitudes and intentions towards helping the victim (dependent variables). This indicates that the independent variable, the social group to which the victim belongs, does not influence bystanders' identification with the victim and bystanders' attitude and intention to provide support. A possible explanation for this might be that the participants identified with the victim not because the victim was a student at a specific University, but because the victim was Dutch. As a result, the participants might have reported to help the victim because they looked similar to the victim based on the Dutch nationality, irrespective of whether the victim belonged to the in-group (Tilburg University) or the out-group (Twente University).

However, as mentioned, the results showed a significant relationship between the mediating variable and the dependent variable, indicating that when bystanders identify with the victim, they are more likely to help this victim. This finding aligns with those obtained by Doosje et al. (1995), who showed that greater feelings of similarity with a person and recognizing a person to be part of the same social category increases the feeling of connectedness with this person. This personal connection leads bystanders to experience more empathetic concern for the victim, which, as a consequence, evokes the motivation to engage in prosocial behavior (Ellemers et al., 1997).

The third hypothesis predicted that bystanders have a more favorable attitude towards

providing support and are more likely to provide support when the victim is anonymous and part of the in-group, compared to when the victim is identifiable and part of the in-group. However, the third hypothesis is not supported as the results were not significant. A possible explanation for this might be the lack of identification with the in-group. As can be seen in Table 4, the differences in helping intentions between in-group and out-group members are minimal, which can be attributed to bystanders having an insufficient connection with Tilburg University or bystanders not considerably disliking members of the University of Twente. Due to the absence of an in-group connection, bystanders may not have shown an increased intention to help anonymous victims of the in-group compared to identifiable victims of the in-group.

Theoretical and practical implications

This study provides one of the first investigations into the influence of the level of anonymity of the victim of online hate speech (anonymous vs. identifiable) on bystanders' attitudes towards helping the victim and their intention to help the victim. Additionally, this study is the first to investigate whether bystanders are more likely to help the anonymous or identifiable victim when they are perceived as belonging to the in-group, compared to the out-group. Although the three hypotheses could not be fully supported, some significant relationships relating to the level of anonymity of the victims were found. Because online hate speech can also be targeted at a group of individuals rather than a specific individual, indicating that there is not always a clear victim, the relevance of investigating victim anonymity in the context of hate speech might be questioned. This research shows that it is relevant to focus on victim anonymity because the findings emphasize the importance of being identifiable as a victim on social media, as bystanders' attitudes towards helping a victim are more favorable when the victim is identifiable rather than anonymous. This favorable attitude towards identifiable victims is caused by feelings of social attraction.

Therefore, social attraction is identified as an important factor in bystanders' decision to provide support to victims. Thereby, this study contributes to research on the social factors that influence positive bystander behavior.

Based on these insights, victims can become aware of the fact that being anonymous on social media can cause bystanders to be less likely to intervene. This awareness is important, as victims can take measures, such as increasing their identifiability on social media, to protect themselves from the potential negative consequences of online hate speech. More specifically, interventions focused on increasing identifiability on social media can be effective in garnering support from bystanders and, in turn, research has found that bystanders' support can reduce the suffering of victims, such as anxiety or depression (Wachs et al., 2022).

Limitations and suggestions for future research

In addition to the theoretical and practical implications, this experimental study is also subject to some limitations. Firstly, in this study, the anonymity of the bystander was not manipulated. However, research investigating the SIDE model often consists of experiments in which both the victim and the bystander's anonymity is manipulated. The SIDE model describes that due to the anonymous environment, bystanders are less aware of their identity and the identity of others (Postmes et al., 1998). As a result, people start to perceive themselves as a group rather than a unique individual. However, as it was not made clear to the bystanders in this study that they were anonymous, they maintained their awareness of their personal identity and may not have shifted their focus to the group identity. Without a focus on the group identity, bystanders may not have perceived themselves as belonging to a group and were, therefore, less likely to engage in prosocial behavior towards members of that group. For this reason, the effects of victim anonymity on bystanders' intention to intervene may have been stronger when the bystanders themselves were also anonymous.

Therefore, a suggestion for future investigations is to manipulate the anonymity of both the victim and the bystander.

Although not significant, there is a trend of bystanders indicating a higher willingness to help anonymous victims who belong to the in-group, compared to identifiable victims who belong to the in-group (see Table 4). This observation is consistent with the findings of the SIDE model. The SIDE model argues that when a victim is anonymous little information about the victim is available, and when these few available cues of the victim's identity imply that the victim and the bystander share a social identity, a strong connection will form (Turner 1984, as cited in Kim & Park, 2011). Contrarily, when a victim is identifiable, it is easier to perceive interpersonal differences, which can make the bystanders feel less connected to the victim (Lee & Nass, 2002). Future research to validate these findings, that takes into account the manipulation of both bystander and victim anonymity, is therefore suggested.

The second limitation relates to the choice of the group to which the hate speech comment was directed. Given that this study focuses on social group membership, deliberate consideration was required to think about which social group would be the target of the hate speech comment. For example, the hate speech comment could not be targeted at Tilburg University students as the participants, who study at Tilburg University, might be likely to offer help not due to their connection with the victim but because they feel personally attacked. Additionally, the hate speech comment could not be targeted at a different nationality or a different age group as the participants, since the victims in the experiment displayed as the in-group then would become a member of the out-group. Therefore, in this study, it was decided to direct the hate speech comment towards Dutch people. However, since all participants were Dutch, the possibility remained that the participants felt personally attacked by the hate speech comment, which could have influenced their responses to the statements regarding their intention to intervene.

The third limitation refers to the manipulation of the social groups. Although the findings of previous research show that people are more likely to help members of the in-group compared to the out-group (Levine et al., 2002), this research has been unable to replicate this finding. This surprising result could be attributed to the in-group, Tilburg University, being too broad or bystanders perceiving the members of the University of Twente not enough as members of the out-group. Therefore, it is suggested that future studies make a clear distinction between in-group and out-group members.

The fourth limitation relates to the sample of this study. In this study, the participants were recruited through convenience sampling, which can lead to over- or underrepresentation of certain groups. This occurred in this study since 71.9% of the participants were female, while in the Dutch population, approximately 50.3% are female (Centraal Bureau voor de Statistiek, 2023). Therefore, the results cannot be generalized to the Dutch population. A representative sample is important, as research on bystander intervention showed that males are more likely to engage in active interventions and females more in passive interventions (Brewster & Tucker, 2015). Since mostly active interventions were measured in this study, lower helping scores were possibly observed as the sample consists mostly of women who are less likely to use this type of intervention. Thus, future research should consider using a purposive sampling method to ensure better generalization of the results.

Lastly, the fifth limitation refers to the use of only one hate speech comment in the experimental design. Various hate speech comments can elicit various emotions in bystanders (Obermaier et al., 2023). For this reason, the responses of the bystanders in this study may not be representative of all different forms of hate speech that can be seen on social media. Therefore, a suggestion for future research is to use a more advanced design by showing multiple hate speech comments to the participants.

Conclusion

The present study aimed to examine whether bystanders between the ages of 18 and 25 years old are more likely to have a favorable attitude and an increased intention to help a victim of online hate speech when the victim is identifiable compared to anonymous. Additionally, the present study examined whether this attitude and intention to help an anonymous or identifiable victim varied based on the social group to which the victim belonged. The results of this study show that bystanders have a more favorable attitude towards helping a victim when the victim is identifiable rather than anonymous. This relationship was mediated by social attraction, indicating that bystanders are more likely to provide support to identifiable victims due to a higher degree of liking. However, since social attraction and social presence were not found to mediate the relationship between the level of anonymity of the victim and the actual helping intentions of bystanders, future research is needed to identify other mediating variables that cause bystanders to feel motivated to help identifiable victims. Lastly, although not significant, the results showed a trend of bystanders being more likely to help anonymous victims of the in-group compared to identifiable victims of the in-group. Therefore, future research should be carried out to better understand the relationship between the level of anonymity of the victim, their social group membership, and bystander intervention. The findings of this study add to the limited amount of research on victim anonymity, by providing insights into the factors that prompt bystanders to engage in positive bystander behavior. This offers victims knowledge of how to present themselves on social media to garner bystander support in situations of online hate speech.

References

Abbott, N., & Cameron, L. (2014). What Makes a Young Assertive Bystander? The Effect of Intergroup Contact, Empathy, Cultural Openness, and In-Group Bias on Assertive Bystander Intervention Intentions. *Journal Of Social Issues*, *70*(1), 167–182. https://doi.org/10.1111/josi.12053

Aizenkot, D. (2020). Social networking and online self-disclosure as predictors of cyberbullying victimization among children and youth. *Children And Youth Services Review*, *119*, 105695. https://doi.org/10.1016/j.childyouth.2020.105695

Benigni, M. C., Joseph, K., & Carley, K. M. (2017). Online extremism and the communities that sustain it: Detecting the ISIS supporting community on Twitter. *PloS One*, *12*(12), e0181405. https://doi.org/10.1371/journal.pone.0181405

Biocca, F., Burgoon, J., Harms, C., & Stoner, M. (2001). Criteria And Scope Conditions For A Theory And Measure Of Social Presence. *Presence: Teleoperators And Virtual Environments*. https://www.researchgate.net/profile/Judee_Burgoon/publication/239665882_Criteria_And_Scope_Conditions_For_A_Theory_And_Measure_Of_Social_Presence/links/53d659f50cf2a7fbb2ea9dc7.pdf

Brewster, M., & Tucker, J. M. (2015). Understanding Bystander Behavior: The Influence of and Interaction Between Bystander Characteristics and Situational Factors. *Victims & Offenders, 11*(3), 455–481. https://doi.org/10.1080/15564886.2015.1009593

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

Brown, A. (2017). What is so special about online (as compared to offline) hate speech? *Ethnicities, 18*(3), 297–326. https://doi.org/10.1177/1468796817709846

Brown, R., Condor, S., Mathews, A. L., Wade, G., & Williams, J. (1986). Explaining intergroup differentiation in an industrial organization. *Journal Of Occupational Psychology*, *59*(4), 273–286. https://doi.org/10.1111/j.2044-8325.1986.tb00230.x

Cascardi, M., Krauss, A., O'Leary, K. D., Loatman, K. L., Sargent, K. S., Grych, J. H., & Jouriles, E. N. (2018). The Bystander Behavior (For Friends) Scale: Factor Structure and Correlation With Prior Victimization. *Journal Of Interpersonal Violence*, *36*(9–10), NP4850–NP4873. https://doi.org/10.1177/0886260518794011

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

Centraal Bureau voor de Statistiek. (2023, 25 augustus). *Mannen en vrouwen*. Centraal Bureau Voor de Statistiek. Geraadpleegd op 18 mei 2024, van https://www.cbs.nl/nl-nl/visualisaties/dashboard-bevolking/mannen-en-vrouwen#:~:text=Zie%20ook%20de%20bevolkingspiramide.,op%20elke%20100%20vrouwen%20zijn.

Chaudhary, M., Saxena, C., & Meng, H. (2021). Countering Online Hate Speech: An NLP perspective. *arXiv (Cornell University)*. https://arxiv.org/pdf/2109.02941.pdf

Corazza, M., Menini, S., Cabrio, E., Tonelli, S., & Villata, S. (2020). A Multilingual Evaluation for Online Hate Speech Detection. *ACM Transactions On Internet Technology*, *20*(2), 1–22. https://doi.org/10.1145/3377323

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

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

DeSmet, A., Veldeman, C., Poels, K., Bastiaensens, S., Van Cleemput, K., Vandebosch, H., & De Bourdeaudhuij, I. (2014). Determinants of Self-Reported Bystander Behavior in Cyberbullying Incidents Amongst Adolescents. *Cyberpsychology, Behavior And Social Networking*, *17*(4), 207–215. https://doi.org/10.1089/cyber.2013.0027

Doosje, B., Ellemers, N., & Spears, R. (1995). Perceived intragroup variability as a function of group status and identification. *Journal of experimental social psychology*, *31*(5), 410-436.

Dovidio, J. F., Gaertner, S. L., Validzic, A., Matoka, K., Johnson, B., & Frazier, S. L. (1997). Extending the Benefits of Recategorization: Evaluations, Self-Disclosure, and Helping. *Journal Of Experimental Social Psychology*, *33*(4), 401–420. https://doi.org/10.1006/jesp.1997.1327

Ellemers, N., Spears, R., & Doosje, B. (1997). Sticking together or falling apart: In-group identification as a psychological determinant of group commitment versus individual mobility. *Journal Of Personality And Social Psychology*, *72*(3), 617–626. https://doi.org/10.1037/0022-3514.72.3.617

Ellemers, N., Spears, R., & Doosje, B. (2002). Self and Social Identity. *Annual Review Of Psychology*, *53*(1), 161–186. https://doi.org/10.1146/annurev.psych.53.100901.135228

Ellison, N. B., Vitak, J., Gray, R., & Lampe, C. (2014). Cultivating Social Resources on Social Network Sites: Facebook Relationship Maintenance Behaviors and Their Role in Social Capital Processes. *Journal Of Computer-mediated Communication*, *19*(4), 855–870. https://doi.org/10.1111/jcc4.12078

Erreygers, S., Pabian, S., Vandebosch, H., & Baillien, E. (2016). Helping behavior among adolescent bystanders of cyberbullying: The role of impulsivity. *Learning And Individual Differences*, *48*, 61–67. https://doi.org/10.1016/j.lindif.2016.03.003

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

Forbes, H. E., Stark, A., Hopkins, S., & Fireman, G. (2019). The Effects of Group Membership on College Students' Social Exclusion of Peers and Bystander Behavior. *The Journal Of Psychology*, *154*(1), 15–37. https://doi.org/10.1080/00223980.2019.1642839

Graf, R. G., & Riddell, J. C. (1972). Helping Behavior as a Function of Interpersonal Perception. *The Journal Of Social Psychology*, *86*(2), 227–231. https://doi.org/10.1080/00224545.1972.9918621

Hawkins, D. L., Pepler, D. J., & Craig, W. M. (2001). Naturalistic Observations of Peer Interventions in Bullying. *Social Development*, *10*(4), 512–527. https://doi.org/10.1111/1467-9507.00178

High, A. C., & Young, R. (2018). Supportive communication from bystanders of cyberbullying: indirect effects and interactions between source and message characteristics. *Journal Of Applied Communication Research*, *46*(1), 28–51. https://doi.org/10.1080/00909882.2017.1412085

Hill, S. E. K., & Courtright, J. A. (1981). Perceived empathy: Its relationship to selected interpersonal variables and student's interpersonal laboratory performance. *Western Journal Of Speech Communication*, *45*(3), 213–226. https://doi.org/10.1080/10570318109374044

Hornsey, M. J. (2008). Social Identity Theory and Self-categorization Theory: A Historical Review. *Social And Personality Psychology Compass*, *2*(1), 204–222. https://doi.org/10.1111/j.1751-9004.2007.00066.x

Jenni, K. E., & Loewenstein, G. (1997). Explaining the Identifiable Victim Effect. *Social Science Research Network*. https://autopapers.ssrn.com/sol3/papers.cfm?abstract_id=5187

Keipi, T., Näsi, M., Oksanen, A., & Räsänen, P. (2017). *Online Hate and Harmful Content: Cross-National Perspectives*. Routledge. https://doi.org/10.4324/9781315628370

Kim, J., & Park, H. S. (2011). The effect of uniform virtual appearance on conformity intention: Social identity model of deindividuation effects and optimal distinctiveness theory. *Computers in Human Behavior*, *27*(3), 1223–1230. https://doi.org/10.1016/j.chb.2011.01.002

Kogut, T., & Ritov, I. (2005). The "identified victim" effect: an identified group, or just a single individual? *Journal Of Behavioral Decision Making*, *18*(3), 157–167. https://doi.org/10.1002/bdm.492

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

Lea, M., & Spears, R. (1992). Paralanguage and social perception in computer-mediated communication. *Journal Of Organizational Computing And Electronic Commerce*, *2*(3–4), 321–341. https://doi.org/10.1080/10919399209540190

Lea, M., Spears, R., & De Groot, D. (2001). Knowing Me, Knowing You: Anonymity Effects on Social Identity Processes within Groups. *Personality & Social Psychology Bulletin*, *27*(5), 526–537. https://doi.org/10.1177/0146167201275002

Lee, E., & Nass, C. (2002). Experimental Tests of Normative Group Influence and Representation Effects in Computer-Mediated Communication. *Human Communication Research*, *28*(3), 349–381. https://doi.org/10.1111/j.1468-2958.2002.tb00812.x

Lee, S., & Feeley, T. H. (2016). The identifiable victim effect: a meta-analytic review. *Social Influence*, *11*(3), 199–215. https://doi.org/10.1080/15534510.2016.1216891

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

Levine, M., Cassidy, C., Brazier, G., & Reicher, S. (2002). Self-Categorization and Bystander Non-intervention: Two Experimental Studies1. *Journal Of Applied Social Psychology*, *32*(7), 1452–1463. https://doi.org/10.1111/j.1559-1816.2002.tb01446.x

Levine, M., Prosser, A., Evans, D. L., & Reicher, S. (2005). Identity and Emergency Intervention: How Social Group Membership and Inclusiveness of Group Boundaries Shape Helping Behavior. *Personality & Social Psychology Bulletin*, *31*(4), 443–453. https://doi.org/10.1177/0146167204271651

Major, B., & O'Brien, L. T. (2005). The Social Psychology of Stigma. *Annual Review Of Psychology*, *56*(1), 393–421. https://doi.org/10.1146/annurev.psych.56.091103.070137

McCroskey, J. C., & McCain, T. A. (1974). The measurement of interpersonal attraction. *Speech Monographs*, *41*(3), 261–266. https://doi.org/10.1080/03637757409375845

Miyazono, K., & Inarimori, K. (2021). Empathy, Altruism, and Group Identification. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.749315

Moreland, R. L., & Zajonc, R. B. (1982). Exposure effects in person perception: Familiarity, similarity, and attraction. *Journal Of Experimental Social Psychology*, *18*(5), 395–415. https://doi.org/10.1016/0022-1031(82)90062-2

Munger, K. (2016). Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment. *Political Behavior*, *39*(3), 629–649. https://doi.org/10.1007/s11109-016-9373-5

Newberry, B. (2001). Raising student social presence in online classes. *WebNet*, 905–910. https://files.eric.ed.gov/fulltext/ED466611.pdf

Obermaier, M. (2022). Youth on standby? Explaining adolescent and young adult bystanders' intervention against online hate speech. *New Media & Society*, 146144482211254. https://doi.org/10.1177/14614448221125417

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

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

Patterson, L. J., Allan, A., & Cross, D. (2016). Adolescent perceptions of bystanders' responses to cyberbullying. *New Media & Society*, *19*(3), 366–383. https://doi.org/10.1177/1461444815606369

Penner, L. A., Dovidio, J. F., Piliavin, J. A., & Schroeder, D. A. (2005). Prosocial Behavior: Multilevel perspectives. *Annual Review Of Psychology*, *56*(1), 365–392. https://doi.org/10.1146/annurev.psych.56.091103.070141

Postmes, T., Spears, R., & Michel, L. (1998). Breaching or Building Social Boundaries? *Communication Research*, *25*(6), 689–715. https://doi.org/10.1177/009365098025006006

Postmes, T., Spears, R., Sakhel, K., & De Groot, D. (2001). Social Influence in Computer-Mediated Communication: The Effects of Anonymity on Group Behavior. *Personality & Social Psychology Bulletin*, *27*(10), 1243–1254. https://doi.org/10.1177/01461672012710001

Räsänen, P., Hawdon, J., Holkeri, E., Keipi, T., Näsi, M., & Oksanen, A. (2016). Targets of Online Hate: Examining Determinants of Victimization Among Young Finnish Facebook Users. *Violence & Victims*, *31*(4), 708–725. https://doi.org/10.1891/0886-6708.vv-d-14-00079

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

Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., & Pierrehumbert, J. B. (2021). HateCheck: Functional Tests for Hate Speech Detection Models. *arXiv*. https://doi.org/10.18653/v1/2021.acl-long.4

Rovira, A., Southern, R., Swapp, D., Campbell, C., Zhang, J. J., Levine, M., & Slater, M. (2021). Bystander affiliation Influences Intervention Behavior: A Virtual Reality study. *SAGE Open*, *11*(3), 215824402110400. https://doi.org/10.1177/21582440211040076

Rusch, H. (2014). The evolutionary interplay of intergroup conflict and altruism in humans: a review of parochial altruism theory and prospects for its extension. *Proceedings Of The Royal Society B: Biological Sciences*, *281*(1794), 20141539. https://doi.org/10.1098/rspb.2014.1539

Salmivalli, C. (2010). Bullying and the peer group: A review. *Aggression And Violent Behavior*, *15*(2), 112–120. https://doi.org/10.1016/j.avb.2009.08.007

Salmivalli, C., Lagerspetz, K., Björkqvist, K., Österman, K., & Kaukiainen, A. (1996). Bullying as a group process: Participant roles and their relations to social status within the group. *Aggressive Behavior*, *22*(1), 1–15.

Sassenberg, K., & Postmes, T. (2002). Cognitive and strategic processes in small groups: Effects of anonymity of the self and anonymity of the group on social influence. *British Journal Of Social Psychology, 41*(3), 463–480. https://doi.org/10.1348/014466602760344313

Schacter, H. L., Greenberg, S., & Juvonen, J. (2016). Who's to blame?: The effects of victim disclosure on bystander reactions to cyberbullying. *Computers in Human Behavior*, *57*, 115–121. https://doi.org/10.1016/j.chb.2015.11.018

Schmid, U. K., Kümpel, A. S., & Rieger, D. (2022). How social media users perceive different forms of online hate speech: A qualitative multi-method study. *New Media & Society*, 146144482210911. https://doi.org/10.1177/14614448221091185

Sellars, A. (2016). Defining hate speech. *Social Science Research Network*. https://doi.org/10.2139/ssrn.2882244

Short, J., Williams, E., & Christie, B. (1976). *The social psychology of telecommunications*. https://ci.nii.ac.jp/ncid/BA22930972

Small, D. A., & Loewenstein, G. (2003). Helping a victim or helping the victim: Altruism and identifiability. *Journal of Risk and uncertainty*, *26*, 5-16.

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

Tokunaga, R. S. (2010). Following you home from school: A critical review and synthesis of research on cyberbullying victimization. *Computers in Human Behavior*, *26*(3), 277–287. https://doi.org/10.1016/j.chb.2009.11.014

Tsang, S., Hui, E. K. P., & Law, B. C. M. (2011). Bystander Position Taking in School Bullying: The Role of Positive Identity, Self-Efficacy, and Self-Determination. *The Scientific World Journal*, *11*, 2278–2286. https://doi.org/10.1100/2011/531474

Tsugawa, S., & Ohsaki, H. (2015). Negative Messages Spread Rapidly and Widely on Social Media. *ACM Conference On Online Social Networks*. https://doi.org/10.1145/2817946.2817962

Wachs, S., Gámez-Guadix, M., & Wright, M. F. (2022). Online Hate Speech Victimization and Depressive Symptoms Among Adolescents: The Protective Role of Resilience. *Cyberpsychology, Behavior, And Social Networking*, *25*(7), 416–423. https://doi.org/10.1089/cyber.2022.0009

Webb, D. J., Green, C. L., & Brashear, T. G. (2000). Development and Validation of Scales to Measure Attitudes Influencing Monetary Donations to Charitable Organizations. *Journal Of The Academy Of Marketing Science*, *28*(2), 299–309. https://doi.org/10.1177/0092070300282010

Weidlich, J., Kreijns, K., Rajagopal, K., & Bastiaens, T. (2018). What Social Presence is, what it isn't, and how to measure it: A work in progress. *EdMedia + Innovate Learning*, 2142–2150. https://www.learntechlib.org/primary/p/184456/

Xu, K., & Lombard, M. (2017). Persuasive computing: Feeling peer pressure from multiple computer agents. *Computers in Human Behavior*, *74*, 152–162. https://doi.org/10.1016/j.chb.2017.04.043

Xu, Q. (2014). Should I trust him? The effects of reviewer profile characteristics on eWOM credibility. *Computers in Human Behavior, 33,* 136–144. https://doi.org/10.1016/j.chb.2014.01.027

Yam, K. C., & Reynolds, S. J. (2014). The Effects of Victim Anonymity on Unethical Behavior. *Journal Of Business Ethics, 136*(1), 13–22. https://doi.org/10.1007/s10551-014-2367-5

You, L., & Lee, Y. (2019). The bystander effect in cyberbullying on social network sites: Anonymity, group size, and intervention intentions. *Telematics And Informatics*, *45*, 101284. https://doi.org/10.1016/j.tele.2019.101284

Zeng, Y., Xiao, J., Li, D., Sun, J., Zhang, Q., Ma, A., Qin, K., Zuo, B., & Liu, X. (2023). The Influence of Victim Self-Disclosure on Bystander Intervention in Cyberbullying. *Behavioral Sciences*, *13*(10), 829. https://doi.org/10.3390/bs13100829

Zeoob. (2022, 9 juli). *ZeOOB: Fake Facebook Post generator with comments*. Zeoob. Geraadpleegd op 15 maart 2024, van https://zeoob.com/generate-facebook-status-post/

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

Appendix A

Experimental stimulus Facebook post

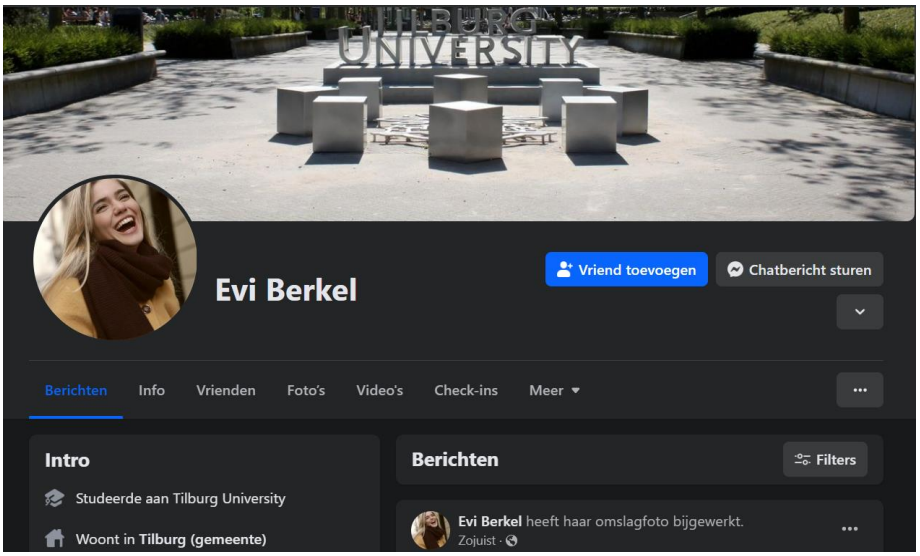
Facebook post of female and male victim together with the hate speech comment

Appendix B

Experimental stimulus Facebook profiles

Figure B1

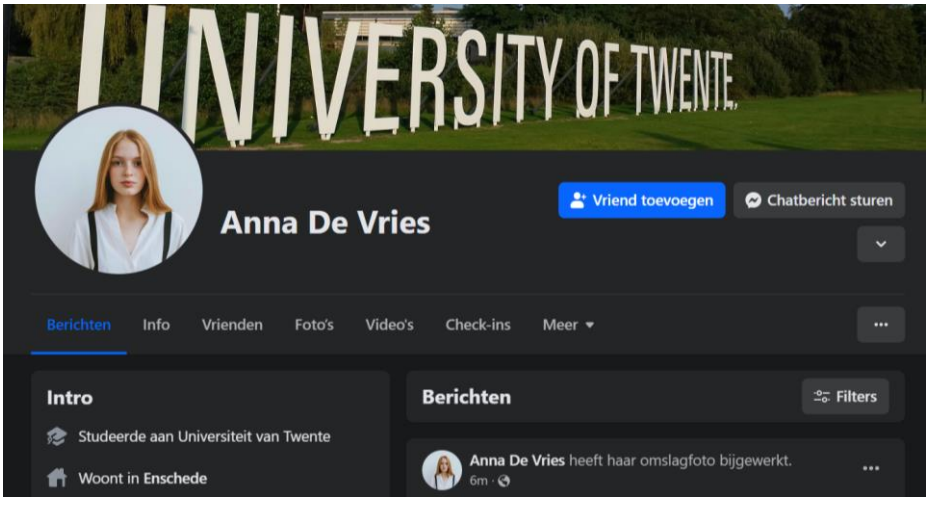
Identifiable victim part of the in-group



Note. The Facebook profile of the male victim of the in-group was displayed in a similar manner, but with a male profile photo.

Figure B2

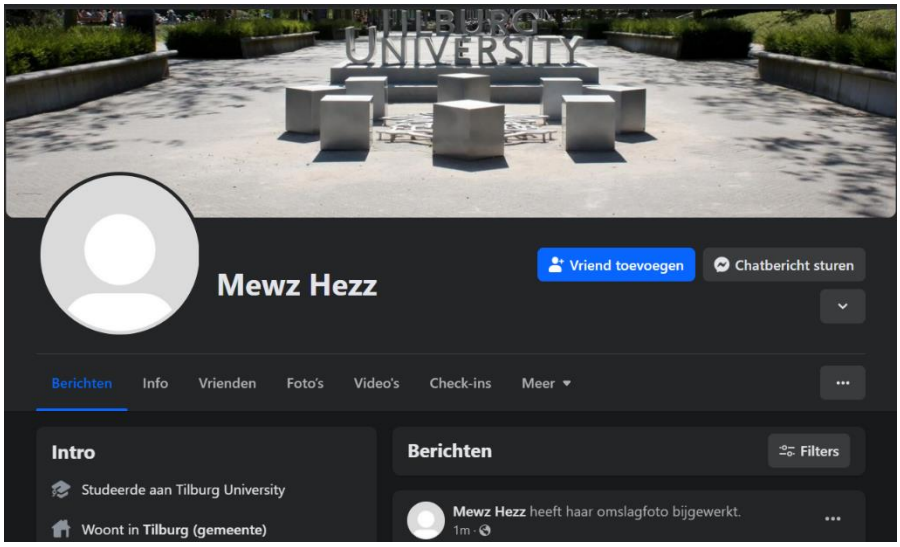
Identifiable victim part of the out-group



Note. The Facebook profile of the male victim of the out-group was displayed in a similar manner, but with a male profile photo.

Figure B3

Anonymous victim part of the in-group



Note. The Facebook profile of the anonymous victim part of the out-group was displayed in a similar manner, but with the University logo of Twente.

Appendix C

Scales used in the questionnaire

The following statement are included in the social attraction scale (McCroskey & McCain, 1974):

1. *I think the victim could be a friend of mine.*

   Ik denk dat het slachtoffer en ik vrienden zouden kunnen zijn.

2. *The victim just would fit into my circle of friends.*

   Ik denk dat het slachtoffer goed in mijn vriendenkring zou passen.

3. *The victim would be pleasant to be with.*

   Ik denk dat het slachtoffer aangenaam is om mee om te gaan.

The following statement are included in the social presence scale (Weidlich et al., 2018):

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

   Het slachtoffer voelde voor mij als een 'echt' persoon.

2. *I can form clear impressions of the victim.*

   Ik kon een duidelijke indruk vormen van het slachtoffer.

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

   Het slachtoffer voelde 'echt' voor mij.

4. *I imagine that I really can see the victim to be in front of me.*

   Ik kon me voorstellen dat ik het slachtoffer voor me zag staan.

5. *It feels as if the victim and I are in close proximity.*

   Het voelde alsof het slachtoffer en ik dicht bij elkaar waren.

6. *I strongly feel the presence of the victim.*

   Ik voelde sterk de aanwezigheid van het slachtoffer.

The following statement are included in the group identification scale (Doosje et al., 1995):

1. *I identify with the victim.*
   Ik identificeer mij met het slachtoffer.
2. *I see myself as the victim.*
   Het slachtoffer lijkt op mij.
3. *I feel a connection with the victim.*
   Ik voel een connectie met het slachtoffer.
4. *I feel strong ties with the victim.*
   Ik voel mij verbonden met het slachtoffer.
5. *The victim and I belong to the same social group.*
   Het slachtoffer en ik behoren tot dezelfde sociale groep.

The following statement are included in the online bystander intervention scale (Obermaier, 2022):

Passive intervention:

1. *I would flag the statements in a comment as 'hate speech'.*

   Ik zou de uitspraken in de comment rapporteren als 'haatspraak'.

2. *I would report the comment.*

   Ik zou de comment rapporteren.

3. *I would like a comment that contradicted the utterances.*

   Ik zou een comment 'liken' die de uitspraken van de pestkop tegenspreekt.

Active intervention:

1. *I would write a comment that factually contradicted the statements.*

   Ik zou een comment schrijven die feitelijk de uitspraken tegenspreekt.

2. *I would refute the statements with facts.*

   Ik zou de uitspraken van de pestkop weerleggen met feiten.

3. *I would write a comment that supported the victim.*

   Ik zou een comment schrijven die het slachtoffer steunt.

4. *I would write a comment that insults the bully.*

   Ik zou een comment schrijven die de pestkop beledigd.

5. *I would confront the bully in a private message.*

   Ik zou de pestkop confronteren in een privébericht.

6. *I would support the victim in a private message.*

   Ik zou het slachtoffer steunen in een privébericht.

The following statement are included in the attitude towards helping others scale (Webb et al., 2000):

1. *Helping the victim is a must.*
   Het helpen van het slachtoffer is een must.

2. *For me, helping the victim is important.*

   Voor mij is het helpen van het slachtoffer belangrijk.

3. *Victims who need help must be helped.*

   Slachtoffers die hulp nodig hebben moeten geholpen worden.

The following statements are included in the group identification scale (Brown et al. 1986) for the manipulation check, depending on the condition the participants were randomly assigned to:

1. *I am a person who sees myself as belonging to Tilburg University (Twente University).*

   Ik voel mij verbonden met Tilburg University (de Universiteit van Twente).

2. *I am a person who identifies with Tilburg University (Twente University).*

   Ik identificeer mij met Tilburg University (de Universiteit van Twente).

3. *I am a person who is glad to belong to Tilburg University (Twente University).*

   Ik ben blij deel uit te maken van Tilburg University (de Universiteit van Twente).

4. *I am a person who feels strong ties with Tilburg University (Twente University).*

   Ik voel een sterke band met Tilburg University (de Universiteit van Twente).

5. *I am a person who considers Tilburg University (Twente University) important.*

   Ik ben iemand die Tilburg University (de Universiteit van Twente) belangrijk vindt.

The following statements are used for the manipulation check of the level of anonymity of the victim (Webb et al., 2000):

1. I could form an impression of the victim.

   Ik kon een impressie vormen van het slachtoffer.

2. I could form an image of the victim.

   Ik kon een beeld vormen van het slachtoffer.

3. The victim was identifiable to me.

   Het slachtoffer was identificeerbaar voor mij.

Appendix D

Factor loadings

Factor loadings of the Principal Component Analysis

Items in the scale	Factors		Uniqueness
	Active intervention	Passive intervention	
I would flag the statements in a comment as 'hate speech'.		.87	.22
I would report the comment.		.87	.19
I would like a comment that contradicted the utterances.	.52	.49	.49
I would write a comment that factually contradicted the statements.	.86		.21
I would refute the statements with facts.	.82		.27
I would write a comment that supported the victim.	.73	.33	.36
I would write a comment that insults the bully.	.86		.26
I would confront the bully in a private message.	.79		.32
I would support the victim in a private message.	.54	.50	.47

Note. The principal component analysis was performed with varimax rotation and eigenvalues over 1. The assumptions for the analysis were met: Bartlett's test of sphericity was statistically significant (χ^2 (36) = 833, $p < .001$) and the KMO measure acquired a reasonably high value (KMO = .86).