SC|M
Studies in Communication and Media

FULL PAPER

# Perceiving threat and feeling responsible

How severity of hate speech, number of bystanders, and prior reactions of others affect bystanders' intention to counterargue against hate speech on Facebook

# Die Gefahr erkennen und sich verantwortlich fühlen

Wie der Schweregrad des Hasskommentars, die Anzahl an Bystandern und vorhergehende Reaktionen anderer die Bereitschaft zu Gegenrede auf Facebook beeinflussen

*Larissa Leonhard, Christina Rueß, Magdalena Obermaier & Carsten Reinemann*

**Larissa Leonhard (M.A.)**, Institute of Communication and Media Studies, University of Leipzig, Burgstr. 21, 04109 Leipzig, Germany; Contact: larissa.leonhard(at)uni-leipzig.de

**Christina Rueß (M.A.)**, Institute of Communication and Media Studies, University of Leipzig, Burgstr. 21, 04109 Leipzig, Germany; Contact: christina.ruess(at)uni-leipzig.de

**Magdalena Obermaier (M.A.)**, Department of Communication Studies and Media Research, LMU Munich, Oettingenstr. 67, 80538 Munich, Germany; Contact: magdalena.obermaier(at)ifkw.lmu.de

**Carsten Reinemann (Prof. Dr.)**, Department of Communication Studies and Media Research, LMU Munich, Oettingenstr. 67, 80538 Munich, Germany; Contact: carsten.reinemann(at)ifkw.lmu.de

FULL PAPER

# Perceiving threat and feeling responsible

## How severity of hate speech, number of bystanders, and prior reactions of others affect bystanders' intention to counterargue against hate speech on Facebook

# Die Gefahr erkennen und sich verantwortlich fühlen

## Wie der Schweregrad des Hasskommentars, die Anzahl an Bystandern und vorhergehende Reaktionen anderer die Bereitschaft zu Gegenrede auf Facebook beeinflussen

*Larissa Leonhard, Christina Rueß, Magdalena Obermaier & Carsten Reinemann*

**Acknowledgement:** The study was developed as part of a master seminar at the Department of Communication Studies and Media Research at LMU Munich. The authors would like to thank Katrin Heimerl and Heinrich-Otto Czeranski for their valuable input and their support in implementing the study.

**Abstract:** Since platform operators are severely challenged to cope with hate speech on social networking sites, countering by individual users is all the more important. Still, it remains unclear to what extent users' intention to actually interfere against hate speech is determined by the context and content of hate speech. Drawing from research on bystander intervention online, we conducted an online experiment ($n$ = 304) to explore the effects of severity of hate speech, number of bystanders, and prior reactions of others on Facebook users' intention to counterargue. Results show that users are less willing to react if the number of bystanders is high, hence providing support for a bystander effect. Also, prior reactions of others lower users' feeling of responsibility to intervene countering hate speech. However, we demonstrate that the severity of hate speech increases users' intention to counterargue if they consider it threatening and concurrently feel responsible to act.

**Keywords:** Counter speech, bystander effect, hate speech, Facebook, experiment

**Zusammenfassung:** Da Plattformbetreiber sozialer Netzwerke wie Facebook noch immer einen angemessenen Umgang mit Hasskommentaren suchen, ist das Eingreifen durch private Nutzer in solchen Situationen unerlässlich. Unklar ist bislang jedoch, inwieweit die Absicht der Nutzer, tatsächlich gegen Hassrede vorzugehen, von deren Inhalt sowie ihrem Kontext abhängt. Ausgehend von der Forschung zur Bystander-Intervention im Online-Bereich führten wir ein Online-Experiment ($n$ = 304) durch, um die Auswirkungen des Schweregrads eines Hasskommentars, der Anzahl an Bystandern, sowie der vorausgegangenen Reaktionen anderer Facebook-Nutzer auf die Absicht des Nutzers, selbst Gegenrede zu tätigen, zu untersuchen. Die Ergebnisse zeigen, dass die Bereitschaft hierzu bei einer hohen Zahl an Bystandern geringer ist, sich also ein Bystander-Effekt beobachten lässt, und vorausgegangene Reaktionen anderer die wahrgenommene Verantwortung der Nutzer, gegen Hasskommentare zu intervenieren, verringern. Zudem wird deutlich, dass der Schweregrad des Hasskommentars die Bereitschaft der Nutzer zu Gegenrede erhöhen kann, und zwar dann, wenn diese den Hasskommentar als bedrohlich empfinden und sich gleichzeitig verantwortlich fühlen, persönlich dagegen tätig zu werden.

**Schlagwörter:** Gegenrede, Bystander-Effekt, Hassrede, Facebook, Experiment

## 1. Introduction

Social networking sites (SNS) are one of the most popular internet applications nowadays, with Facebook being by far the most widespread SNS worldwide (Newman, Fletcher, Kalogeropoulos, Levy, & Nielsen, 2017, p. 11). By commenting, sharing, or rating content, the platform potentially constitutes an open space for private as well as for public discourse. However, recent developments paint a somewhat disillusioning picture. In fact, SNS like Facebook are increasingly under fire for not sufficiently restricting the dissemination of so-called *hate speech* aimed at devaluating others "because of their religion, race, ethnicity, gender, sexual orientation, national origin, or some other characteristic that defines a group" (Hawdon, Oksanen, & Räsänen, 2017, p. 254).

A recent study in Germany shows that in 2016, two-thirds of internet users had already been confronted with hateful messages online (Kaspar, Gräßer, & Riffi, 2017, p. 9). According to an online survey from the US, nearly half of the content perceived as hateful by the participants referred to race or ethnicity (Costello, Hawdon, Ratliff, & Grantham, 2016), with this topic also being the major subject of a surge in hate speech across Europe triggered by the so-called refugee crisis starting in 2015 (Ross et al., 2016). For individuals being the target of hate speech, it can have severe consequences ranging from short-term emotional reactions like shock, loneliness, or anger to long-term behavioral effects like increased mistrust in contact with strangers or social exclusion (Boeckmann & Liew, 2002; Leets, 2002; Obermaier, Hofbauer, & Reinemann in this special issue). In addition to these negative effects for individuals targeted, hate speech can also lead to undesirable effects at the societal level. For one, uninvolved witnesses of such statements are at risk of perceiving the climate of opinion distorted in the direction of the views expressed in the hate speech, which can reduce their willingness to speak out against those statements (Zerback & Fawzi, 2017). For another, frequent contact with hateful content can lower the inhibition threshold for hateful countering or further hate speech, and, in turn, can exacerbate social polarization tendencies (Leets & Giles, 1997). Yet, the proliferation of such hateful content is facilitated by specific characteristics of online communication, resulting in what Suler (2004) describes as online disinhibition effect. Although Facebook requires the use of a clear name, it can be assumed that many of its users bypass this requirement by using a name that sounds real but is not their own, thereby masking their identity. But even though most users sign in with their real names, online platforms like Facebook enable communications and actions that are disengaged from users' 'real' life. That means users can only know as much about another user as she discloses online about her own identity, so that even using a real name might not imply opening up about one's actual personal lifestyle and character. Moreover, an online disinhibition may not just be amplified due to users' anonymity, but also due to their mutual invisibility, meaning that they cannot see or hear their interlocutors and, hence, are unable to decode dialogue partners' body language or facial expressions. This lack of social and contextual cues as well as the mutual invisibility of dialogue partners in the online context can thus tempt users to express themselves in more extreme ways than they would in a face-to-face interaction (Spears & Lea, 1992). Hence, the non-visibility of the effects hate speech has on a victim may contribute to a lower inhibition threshold. However, both legal sanctions against hate speech and censorship by platform operators raise problematic issues in terms of freedom of expression and public discourse (Benesch, 2014) – a situation perfectly illustrated in Germany by the current debate about a law to improve the enforcement of speech regulation in social networks (Act to Improve Enforcement of the Law in Social Networks [Network Enforcement Act], 2017). Most of the time, the line between censorship and necessary legal interdiction of hatred is narrow. For this reason and, additionally, since passive behavior of bystanders may be perceived as implicit approval of hate speech, reactions of individual users such as countering are all the more important in dealing with the problem of hate speech (e.g., "Online Civil Courage Initiative," n.d.; Schieb & Preuss, 2016). Countering can be understood as "common, crowd-sourced response to extremism or hateful content . . . capable of dealing with extremism from anywhere and in any language" while maintaining "the principle of free and open public spaces for debate" (Bartlett & Krasodomski-Jones, 2015, p. 5).

The huge and rapid proliferation of hate speech online and the relative lack of research on factors influencing both the origins and proliferation of hate and counter speech make these highly relevant research topics. In one of the few existing studies, Schieb and Preuss (2016, and in this special issue) identify the proportion of SNS users supporting the hate speech, the susceptibility of bystanders to the influence of counter speakers, the timing of the counter speech, and the extremity of its position as decisive factors for successful countering. Ziegele, Jost, Bormann, and Heinbach (in this special issue) illustrate how different moderation strategies applied by journalists on Facebook pages of German news outlets vary in success with regard to the containment of uncivil comments. In addition to factors concerning the *context* of hate speech, the severity of its *content* could also be crucial in determining whether bystanders are willing to perform countering. Against this backdrop, this study inquires which factors influence bystanders' intention to counter hate speech online. More precisely, we ask whether and to what extent severity of hate speech, number of bystanders, and prior reactions of others influence the intention to counterargue hate speech on Facebook. In order to shed light on this question, we conducted an online experiment, varying the content of a fictitious hate speech against asylum seekers, the number of bystanders who had already seen the hateful comment as well as prior reactions of other users. As research on bystanders' willingness to counter hate speech on SNS is quite scarce, our study, contributes to a better understanding of this highly relevant topic.

## 2. Influences on bystanders' willingness to intervene in hate speech

In this paper, we argue that encounters with hate speech can be understood as a specific form of an emergency and, hence, counterarguing can be conceptualized as a specific form of helping behavior by bystanders. This allows us to borrow from research into helping behavior when looking for reasons to engage in counterarguing or not. Generally speaking, to intervene in an emergency, bystanders need to (1) notice a critical situation, (2) be aware of the fact that the situation is an emergency (to a certain degree), (3) consider themselves personally responsible to intervene, (4) reflect how to help, and (5) decide to intervene and to implement that decision (Latané & Darley, 1970). These stages of the decision-making process for helping behavior have been investigated in experimental settings for different types of helping behavior and in various kinds of emergencies (for an overview, see Fischer et al., 2011; Latané & Nida, 1981).

An emergency of this sort may stem, for example, from antisocial behavior of individuals offline as well as in the online environment. Cyberbullying is one form of antisocial online behavior for which bystander reactions already have been investigated quite extensively. Several studies could demonstrate the above-mentioned five stages to affect bystander intervention even in the online context (Obermaier, Fawzi, & Koch, 2016; Weber, Ziegele, & Schnauber, 2013). Surely, hate speech differs from cyberbullying in some aspects (e.g., see Hawdon et al., 2017). For instance, while cyberbullying targets individuals (even though a whole group of individuals may feel victimized), hate speech explicitly displays hostility towards a whole group, for the most part a minority, precisely because of their collective identity (e.g., in terms of origin, sexual orientation, etc.). Hate speech can also differ from cyberbullying concerning its time span: Cyberbullying is defined as harassment of another individual that takes place repeatedly and for a longer period of time (Tokunaga, 2010), whereas hate speech may consist of one single incident. Yet, due to the high range and the possibilities to easily disseminate and store content online, even a single incident of harassment or hate speech can potentially lead to repeated victimization (e.g., Obermaier, Fawzi, & Koch, 2015). Also, both forms of antisocial behavior are based on an imbalance of power between victims and offenders and are comparable in terms of their goal to degrade another individual by means of (non-)verbal attacks. Hence, we believe that the phenomena of cyberbullying and hate speech have sufficient similarities to assume that bystanders in an incident of hate speech also have to complete the decision-making process proposed by the model (Latané & Darley, 1970) to intervene. In the following, we focus on the first three steps of the model as well as on the intention to intervene in an incident of hate speech as dependent variable. More precisely, we tested the sequence of severity of hate speech (step 1), perception of threat (step 2), feeling of personal responsibility (step 3), and intention to counterargue (step 5), as it has proven to be a robust model in research on bystander intervention in antisocial behavior (Fischer et al., 2011; Obermaier et al., 2016).

## 3. Content-related influences on bystanders' willingness to counterargue

As stated by Latané and Darley (1970), noticing a critical situation and recognizing it as an emergency are the fundamental first steps necessary for intervention. Such an assessment should be easier the more obvious the respective emergency is. In line with research on bystander intervention in cyberbullying and racist comments, we therefore assume that the severity of an incident of antisocial behavior increases bystanders' intention to intervene (Dickter & Newton, 2013; Fischer et al., 2011; Obermaier et al., 2016). Thus, a higher degree of threat and harm related to an incident of hate speech online should result in a higher probability of intervention on part of users witnessing the incident. Therefore, for hate speech explicitly calling for violence against a group of victims, countering should be more likely than for hate speech containing mere insults against that group. Hence, we formulate the following hypothesis (Figure 1):

*H1a: The more severe an incident of hate speech is, the more likely an individual intends to counterargue.*

As included in the second step of Latané and Darley's (1970) model, recognizing the situation as an emergency is a crucial step in the decision-making process for bystander intervention. However, the evaluation of the degree of threat to a situation is by definition subjective. Therefore, for an individual to take action, it is necessary that she perceives the situation at hand as an emergency. Thus, especially if a user categorizes the hate speech as a high threat to the victim group, she will be more willing to intervene. Therefore, we assume:

*H1b: The severity of an incident of hate speech will positively affect an individual's intention to counterargue in case the situation is perceived as threatening.*

The perception of a situation as an emergency, however, is not yet a guarantee that helping behavior will actually take place. Rather, it is necessary that a bystander feels personally responsible for intervening in a situation perceived as an emergency (Dickter & Newton, 2013; Latané & Darley, 1970; Markey, 2000). Hence, recognizing hate speech as a threatening incident has to be followed by the important step of assuming personal responsibility to intervene for bystanders to engage in countering. Therefore, the more severe hate speech is and the more it is perceived as a threat, the greater the feeling of a personal responsibility to intervene should be. Accordingly, we hypothesize:

*H1c: The severity of an incident of hate speech will positively affect an individual's intention to counterargue mediated by the perception of the situation as threatening and, as a consequence, by an increased feeling of personal responsibility.*

## 4. Context-related influences on bystanders' willingness to counterargue

Beyond the content of hate speech, features of the context in which it is situated could be decisive for bystanders' willingness to intervene as well. On SNS like Facebook, key characteristics of the situational context are, for example, the number of other users who have already seen the hate speech as well as the reactions of others to the hate speech such as commenting it in support of the hater or the group victimized.

### 4.1 Impact of the number of bystanders on intention to counterargue

The group dynamic of bystanders being aware of a situation where individuals need help has been extensively researched, resulting in the observation of the so-called bystander effect: Accordingly, the feeling of personal responsibility decreases with the number of bystanders being present, resulting in less willingness to intervene (Fischer et al., 2011; Latané & Darley, 1970; Obermaier et al., 2016). Although the effect originally states that the *physical* presence of others hampers bystander intervention (Darley & Latané, 1968; Latané & Darley, 1968; Latané & Nida, 1981), several studies indicate its validity also for incidents in computer-mediated communication like cyberbullying (Bastiaensens et al., 2014; Blair, Foster Thompson, & Wuensch, 2005; Markey, 2000; Palasinski, 2012). That is, if a larger number of bystanders is witnessing an incident of cyberbullying, a bystander intervention is less likely than if only a few bystanders are present. Considering the potential visibility of hate speech on Facebook to the public or at least to part of the Facebook community (e.g., members of respective groups on Facebook), a potentially high number of other users could become aware of a hate speech incident. Hence, a bystander effect might occur in an incident of hate speech, too. One could further assume that a high number of bystanders affects the perceived urgency also by means of affecting assumptions of how recently an event happened. More precisely, if several thousand users saw an incident of hate speech, the event might be regarded as less recent compared to hate speech with only a handful of witnesses, as it usually takes time for a post to reach such a large number of users. Consequently, the hate speech incident could be interpreted as having lost of its topicality, which might render any (further) intervention unnecessary. Therefore, we suppose (Figure 1):

*H2a: The higher the number of bystanders to an incident of hate speech is, the less likely an individual intends to counterargue.*

According to Latané and Darley (1970), this relation between the number of bystanders and the decision to intervene is mediated by the feeling of personal responsibility. Specifically, they assume that the number of bystanders affects intention to intervene due to so-called *diffusion of responsibility*: Individuals witnessing a situation of emergency tend to mentally divide the responsibility to intervene among all the bystanders watching. As a consequence, they themselves perceive to be less responsible to intervene the more others are also witnessing the incident. In the context of computer-mediated communication in general and on

SNS specifically, there is already some evidence demonstrating such an indirect bystander effect to occur in antisocial incidents (Obermaier et al., 2016). Hence, we assume:

*H2b: The number of bystanders to an incident of hate speech will negatively affect an individual's intention to counterargue mediated by a decreased feeling of personal responsibility.*

### 4.2 Impact of prior reactions to hate speech on intention to counterargue

Furthermore, bystanders' intention to intervene can be affected not only by the (physical or virtual) presence of others, but also by their action taken. In the case of hate speech on SNS, such reactions can take various forms differing in the level of directness with reporting a post to Facebook as a violation of its terms of use being an indirect form and commenting, liking, or adding an emotion being a direct form of reaction since the author of the hate post would be able to see it. Also, these indirect or direct reactions can be either supportive or opposing (Ernst et al., 2017). Supportive reactions on Facebook, on the one hand, include liking and commenting the hate speech in a confirmative or even reinforcing manner. On the other hand, countering reactions can range from reporting it to Facebook as an indirect form right up to adding an anger-smiley or commenting the hate speech in a dissenting way as direct forms. Most frequently, both supporting and opposing reactions occur in the context of one and the same hate post.

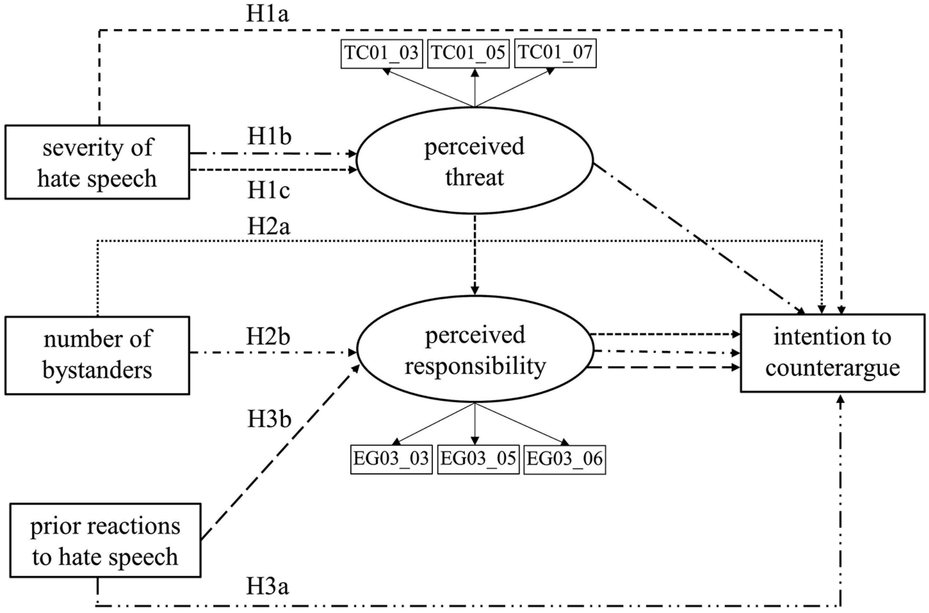
According to spiral of silence theory (Noelle-Neumann, 1974), the willingness of people to speak out publicly depends on their perceptions of majority opinions. There is some evidence that a decrease in the willingness to speak out when perceiving the own opinion to be in the minority also occurs online, for instance, in online discussion groups or comment sections (Nekmat & Gonzenbach, 2013; Woong Yun & Park, 2011). Moreover, building on spiral of silence as well as exemplification theory there are some preliminary findings that support for or mixed reactions to an antisocial incident online can diminish an individual's intention to intervene, whereas countering reactions of others could boost it (Zerback & Fawzi, 2017). Thus, concerning hate speech on Facebook, reactions of others might serve as a benchmark for assessing the climate of opinion regarding the topic of the hate speech. As a consequence, with reactions of others unanimously rejecting the hateful utterance, willingness to counter the hate speech should be stronger compared to a situation where there are mixed reactions or even support for the hater. Hence, we hypothesize:

*H3a: Unanimously countering reactions to hate speech of others will more strongly increase an individual's intention to counterargue compared to mixed reactions (countering and supporting hate speech) or no reaction at all.*

On the contrary, in a hate speech situation where neither countering nor mixed reactions are present, bystanders could be more willing to intervene as a matter of particular urgency. Hence, witnessing an incident of hate speech and perceiving that nobody has intervened yet in order to support the victim could boost users feeling responsible to engage in countering the hate speech (Latané & Darley, 1970). Therefore, mediated by an increased feeling of responsibility, perceiving no reactions to the hate speech compared to already having witnessed reactions of others could increase the intention to intervene on the contrary. Consequently, we suppose:

*H3b: No reactions to hate speech of others will increase an individual's intention to counterargue compared to already existing (countering or mixed) reactions mediated by an increased feeling of personal responsibility.*

**Figure 1: Proclaimed model on the effects of severity of hate speech, number of bystanders, and prior reactions to hate speech on the intention to counterargue**



## 5. Method

### 5.1 Design and participants

To test our hypotheses, we conducted an online experiment in a 2 x 2 x 3 between-participants design. Participants were recruited via a non-commercial online access panel that, although lacking representativeness, provides more heterogeneity than student convenience samples (Leiner, 2016). In total, 556 Facebook users took part in the survey and were randomly assigned to the experimental groups. We excluded participants who did not (correctly) recall the number of bystanders.[1] Also, we excluded participants who did not recall the kind of reactions to the incident of hate speech correctly, that is, depending on the experimental condition, if other users had commented below the hate post or not. In addition, participants who sympathized with the hater have not been taken into account, as their willingness to comment could be interpreted as intent to endorse the hater instead of as intent to engage in counter speech. After that, 304 participants remained in the sample (64% female, age: $M$ = 32 years, $SD$ = 12.3, political predisposition[2]: $M$ = 34.10, $SD$ = 16.94), whereby each of the twelve experimental groups comprised on average 25 participants ($SD$ = 3.6). The composition of the experimental groups did not significantly differ regarding gender, $\chi^2(22)$ = 27.41, $p$ = .16, age, $F(11, 292)$ = .62, $p$ = .81, and political predisposition, $F(11, 286)$ = 1.15, $p$ = .32.

## 5.2 Procedure and independent variables

Participants were asked to read a screenshot of a (fictitious) post on Facebook offending asylum seekers in Germany. Participants were instructed to imagine being a member of the Facebook group in which the post was released, but not knowing any of its members personally. Also, they were told to take a close look at the post and its attributes such as the comments beneath it (in the respective conditions). The post representing hate speech was directed against asylum seekers. More specifically, asylum seekers were severely insulted several times (e.g., “filthy bunch”) and the hater requested that refugees leave the country (Table 1). The post was created in the design of a Facebook thread. However, the profile pictures as well as the names of the authors of the post and of the comments were blurred. We manipulated our three independent variables as follows. As a first factor, we varied the *severity of hate speech*: The (fictitious) post in the condition of medium severe hate speech contained insults against refugees (e.g., “They’re all lazy suckers”) and an indirect statement for them to leave the country (“They don’t belong here!!”). In the highly severe version of hate speech the wording of the insults was more hostile including dehumanization (e.g., “Throw that vermin out of our country”) and explicitly inciting to violence against asylum seekers in order to make them leave the country (“We have to fight!!!!!!”, Table 1). The second factor was the *number of bystanders* that is the number of users who had seen the post beforehand (indicated in the screenshot by “seen by …”). Based on findings from Blair et al. (2005) and Obermaier et al. (2016) according to which differences in the willingness to intervene only emerge between very low and very high numbers of by-

[1] In order to assess the correct recall of the number of bystanders, participants were asked to indicate, by means of an open entry, how many users had already seen the post, accompanied by an instruction to make a rough estimate in case they could not remember the exact number. Participants in the few-bystander-condition were excluded if their answer was 20 and higher, participants in the many-bystander-condition were excluded if their answer was below 1000 or higher than 6000. This approach was based on findings from previous studies and applied in order to provide comparability of results (e.g., see Obermaier et al., 2016; You & Lee, 2018).

[2] Political predisposition was assessed using a slider, with the leftmost position corresponding to the value 0 and the rightmost position corresponding to the value 100.

standers, we indicated that the post has already been “seen by 4” vs. “seen by 3,997” group members. As a third factor, we varied *prior reactions to hate speech*: There was either no reaction or a comment presenting counter speech, that is rejecting the statement of the hater, combined with a comment of a third user either positively (i.e., unanimously countering the hate speech) or negatively reacting to the counter speech (i.e., mixed reactions to the hate speech, Table 1).

**Table 1.** Wording of stimuli

| **Stimulus example (version: medium severe hate speech, unanimously countering)** | | **Stimulus example (version: highly severe hate speech, mixed reactions)** |
|---|---|---|
| “We don’t want this filthy bunch of refugees in our country!!! They’re all lazy suckers. They don’t belong here!!” | **Hate speech post (user A)** | “This filthy bunch of refugees!!! Stop it!!!!!! Throw that vermin out of our country… and send it right back to where they came from… And shoot every bastard that comes crawling back here. We have to fight!!!!!!” |
| “That’s nonsense! A little more affection and something like compassion and humanity wouldn’t hurt you …” | **Countering (user B)** | “That’s nonsense! A little more affection and something like compassion and humanity wouldn’t hurt you …” |
| “That’s how I see it, too. Refugees are welcome here!” | **Reaction to countering (user C)** | “Not for them! Refugees are NOT welcome here!” |

*Note.* As the questionnaire was administered in German, all wordings represent English translations of the original stimulus versions.

### 5.3 Dependent measures

The treatment check consisted of questions on whether the hate speech contained an *incitement to violence* since that represents the key difference between the medium and the highly severe hate speech (0 = “no,” 1 = “yes,” 2 = “I do not recall”).[3] Also, participants had to *roughly assess the number of bystanders* having already seen the post (7-point scale, 1 = “very few,” 7 = “very many”) and recall the *exact number of bystanders*. In addition, participants had to indicate whether other users have already commented the post or not, and if the users all agreed with the hater in the respective experimental condition.

We measured the *perceived threat* of the hate speech by asking to what degree it is “threatening,” “harmful to the affected people,” and “has the potential to incite to violence” (7-point scales, 1 = “I completely disagree,” 7 = “I completely agree,” $M$ = 5.99, $SD$ = 1.23, $\alpha$ = .82). The *perceived personal responsibility* to intervene was assessed as follows: “It is my personal responsibility to intervene,” “It is my job to intervene,” and “It is my duty to take action” (7-point scales, 1 =

[3] As the questionnaire was administered in German all following items represent English translations of the original items.

"not at all important," 7 = "very important," $M = 3.88$, $SD = 2.05$, $\alpha = .94$). We inquired the *intention to counterargue* asking for the intention to comment against the hate speech (7-point scale, 1 = "highly unlikely," 7 = "highly likely," $M = 2.09$, $SD = 1.70$).

## 6. Results

### 6.1 Treatment check

Before testing the hypotheses, we conducted a treatment check. The majority of the remaining 304 participants (72%) correctly recalled the existence respectively non-existence of an incitement to violence in the post representing hate speech, while 21 percent gave the wrong answer and seven percent stated not being able to recall, $\chi^2(2) = 93.95$, $p < .001$. With regard to the number of bystanders, we have already ensured that the discrepancies between the correct number and participants' estimates remain modest, but we further tested the subjective assessment of the number of bystanders. Indeed, 3,998 bystanders were perceived as a significantly higher number of present users ($M = 5.70$, $SD = 1.29$) than 4 bystanders ($M = 1.98$, $SD = 1.21$), $t(298.31) = -25.92$, $p < .001$. In terms of prior reactions to the hate speech, all remaining participants correctly recalled that the three users disagreed in so far as the hate post was accompanied by two comments. Therefore, according to the treatment check, our experimental manipulation was successful.

### 6.2 Direct effects of severity of hate speech, number of bystanders, and reactions of others on intention to counterargue, perceived threat, and feeling of responsibility

Before testing the presumed model, we checked the direct effects of our treatment. In a first step, we investigated the proposed direct effects of severity of hate speech (*H1a*), number of bystanders (*H2a*), and prior reactions of other users (*H3a*) on participants' intention to counterargue. Hence, we conducted an analysis of variance using the severity of hate speech, the number of bystanders, and prior reactions of other users as independent variables. First, contrary to our assumption, we found no significant main effect of severity of hate speech on users' willingness to counterargue, $F(1, 292) = .07$, $p = .79$, $\eta^2_{part} < .001$. Hence, participants did not indicate to be more willing to counterargue if the hate speech contained an incitement to violence ($M = 2.07$, $SD = 1.69$) than if it was less severe ($M = 2.10$, $SD = 1.71$). Thus, *H1a* is rejected.

Second, if a high number of bystanders had already seen the Facebook post, participants' intention to counterargue was significantly lower ($M = 1.82$, $SD = 1.37$) than with only few witnesses of the incident ($M = 2.40$, $SD = 1.98$), $F(1, 292) = 8.48$, $p = .004$, $\eta^2_{part} = .03$. This result provides support for a direct bystander effect when being confronted with hate speech online. However, a rather weak effect emerged which could also be due to the overall low levels of willingness to intervene in the sample. Thus, *H2a* is preliminarily supported. Third, there was no significant main effect of prior reactions, $F(2, 292) = .06$, $p = .95$, $\eta^2_{part} < .001$. Participants were not more willing to intervene if they read a countering along with a hate comment ($M = 2.13$, $SD = 1.72$), if two users had already countered the hate speech ($M = 2.07$, $SD = 1.64$), or if no one had reacted yet ($M = 2.06$, $SD = 1.76$). Therefore, *H3a* is rejected. Also, no significant interaction effects between severity of hate speech and number of bystanders, $F(1, 292) = .66$, $p = .42$, $\eta^2_{part} = .002$, severity of hate speech and prior reactions, $F(2, 292) = .54$, $p = .58$, $\eta^2_{part} = .004$, and number of bystanders and prior reactions emerged, $F(2, 292) = .63$, $p = .53$, $\eta^2_{part} = .004$. Also, there was no threefold interaction, $F(2, 292) = .05$, $p = .95$, $\eta^2_{part} < .001$.

In a second step, we scrutinized how the treatment was related to direct dependent variables proposed in the model, namely the perceived threat and the feeling of responsibility. First, we computed an analysis of variance on perceived threat using severity of hate speech, number of bystanders as well as prior reactions as independent variables. The analysis demonstrated a main effect of the severity of hate speech, $F(1, 292) = 4.99$, $p = .03$, $\eta^2_{part} = .02$. Participants regarded the post with incitement to violence as very threatening ($M = 6.15$, $SD = 1.23$) and significantly more so than the less severe hate speech ($M = 5.82$, $SD = 1.22$). Also, participants categorized the less severe hate speech as a serious threat, too, since the mean value is above the scale midpoint. However, neither a main effect of number of bystanders on perceived threat, $F(1, 292) = .49$, $p = .48$, $\eta^2_{part} = .002$, nor of prior reactions appeared, $F(2, 292) = 2.81$, $p = .06$, $\eta^2_{part} = .02$. Moreover, neither an interaction between degree of severity and number of bystanders, $F(1, 292) = 2.81$, $p = .06$, $\eta^2_{part} = .02$, degree of severity and prior reactions, $F(2, 292) = 1.13$, $p = .32$, $\eta^2_{part} = .01$, nor number of bystanders and prior reactions could be observed, $F(2, 292) = .42$, $p = .66$, $\eta^2_{part} = .003$. Also, no threefold interaction emerged, $F(2, 292) = .33$, $p = .72$, $\eta^2_{part} = .002$.

Second, we conducted another analysis of variance with the feeling of responsibility as dependent variable and number of bystanders, prior reactions of others, and severity of hate speech as independent variables. However, the analysis yielded no main effect of the number of bystanders: If the number of bystanders was high, participants did not feel less responsible to act ($M = 3.78$, $SD = 2.04$) than if there were only a few bystanders ($M = 4.00$, $SD = 2.07$), $F(1, 292) = .75$, $p = .39$, $\eta^2_{part} = .003$. Also, we found no significant main effect of prior reactions on feeling of personal responsibility, $F(2, 292) = .54$, $p = .58$, $\eta^2_{part} = .004$. Hence, there was no difference in terms of feeling of personal responsibility to intervene if there was no comment ($M = 4.06$, $SD = 2.08$) compared to two countering comments ($M = 3.81$, $SD = 1.97$), or one countering and one hate comment ($M = 3.76$, $SD = 2.13$). The analysis further demonstrated a weak significant effect of severity of hate speech, $F(1, 292) = 5.80$, $p = .02$, $\eta^2_{part} = .02$, with participants feeling more personally responsible to intervene if the post contained an incitement to violence ($M = 4.17$, $SD = 2.08$) than if it was less severe ($M = 3.59$, $SD = 1.99$). There was neither a significant interaction between number of bystanders and prior reactions, $F(2, 292) = 1.38$, $p = .25$, $\eta^2_{part} = .01$, between number of bystanders and degree of severity, $F(1, 292) = .06$, $p = .80$, $\eta^2_{part} < .001$, nor between degree of severity and prior reactions, $F(2, 292) = 2.43$, $p = .09$,

$\eta^2_{part}$ = .02. Also, no threefold interaction was evident, $F(2, 292) = .60$, $p = .55$, $\eta^2_{part}$ = .004.

## 6.3 Indirect effects of severity of hate speech, number of bystanders, and prior reactions on intention to counterargue

In order to test the indirect effects, we used structural equation modeling (SEM) with maximum likelihood estimation (ML)[4] utilizing Mplus (Muthén & Muthén, 2010). First, we specified a measurement model including the six items which were assumed to represent the latent variables *perceived threat* and *perceived responsibility*, respectively. The model showed a very good fit for the data (Hu & Bentler, 1999): $\chi^2(8) = 7.84$, $p = .45$; CFI = 1.00, TLI = 1.00; RMSEA = .00, $p$ = .78; SRMR = .02. Hence, the included items provide a highly reliable reflection of the respective factors.[5] Subsequently, this measurement model was integrated in the following structural model: We included the dichotomous variables for severity of hate speech (0 = medium severity, 1 = high severity) and for number of bystanders (0 = 4, 1 = 3,997) as predictors in the structural equation model (Figure 2). As a third predictor, we dummy coded the variable for prior reactions to hate speech (0 = no reaction, 1 = countering/mixed reactions), due to the small difference in mean values of perceived responsibility if no other user had commented, or if there were two countering, or one countering and one hate comment. Intention to counterargue served as dependent variable, while the latent variables (perceived threat and perceived responsibility) were modeled as mediators of the predictors' effects on intention to counterargue. Table 2 shows zero-order correlations among all variables included.

**Table 2. Zero-order correlations between included variables**

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1. Severity of hate speech | 1 | | | | | |
| 2. Number of bystanders | –.03 | 1 | | | | |
| 3. Prior reactions to hate speech | –.01 | .03 | 1 | | | |
| 4. Perceived threat (factor) | .13* | –.04 | .12*a | 1 | | |
| 5. Perceived responsibility (factor) | .14*a | –.05 | –.06 | .39*** | 1 | |
| 6. Intention to counterargue | –.01 | –.17** | .01 | .21*** | .44*** | 1 |

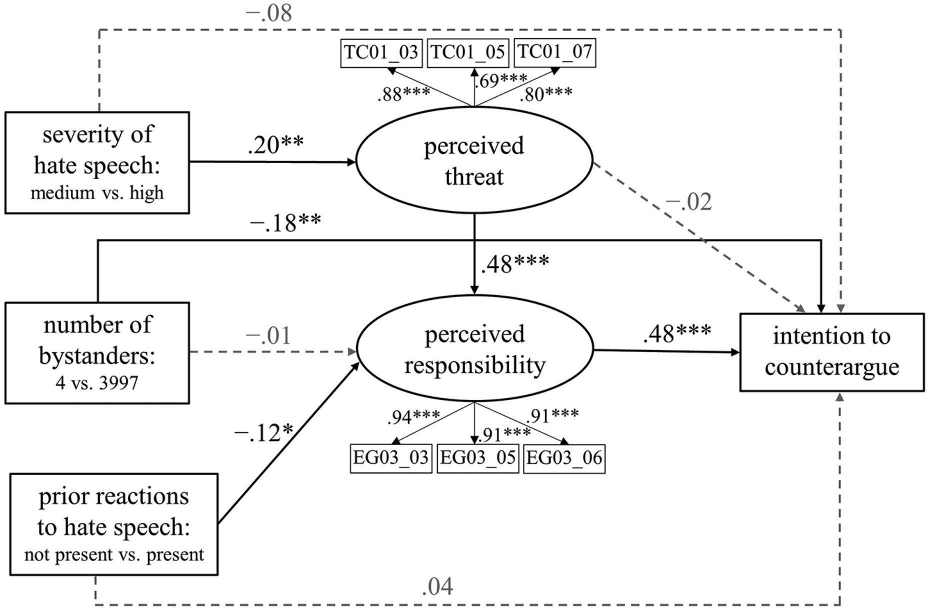
*Note*. Bivariate Pearson correlation coefficients. $n = 304$. * $p < .05$; ** $p < .01$; *** $p < .001$.
a. These paths, though significant, were not added to the structural equation model since we aimed to test the theoretical model in the narrowest possible form.

[4] We chose Maximum Likelihood as estimation method because it is quite robust even if there are moderate departures from assumptions of normal distribution (Bagozzi & Yi, 2012, p. 29).

[5] Results of the confirmatory factor analysis (CFA): factor loadings for *perceived threat*, "threatening": $\lambda = .87$, $p < .001$; "harmful to the affected people": $\lambda = .70$, $p < .001$; "has the potential to incite to violence": $\lambda = .81$, $p < .001$. Factor loadings for *perceived responsibility*, "It is my personal responsibility to intervene": $\lambda = .94$, $p < .001$; "It is my job to intervene": $\lambda = .90$, $p < .001$; "It is my duty to take action": $\lambda = .91$, $p < .001$.

The model fit indices indicate a good fit (Hu & Bentler, 1999): $\chi^2(27) = 29.39$, $p = .34$; CFI = 1.00, TLI = 1.00; RMSEA = .02, $p = .90$; SRMR = .04. It can be concluded that the model adequately describes the empirical data. The full model is shown in Figure 2.

**Figure 2: Model on the effects of severity of hate speech, number of bystanders, and prior reactions to hate speech on the intention to counterargue**



*Note:* Standardized path coefficients. $n = 232$. $\chi^2(27) = 29.39$, $p = .34$; CFI = 1.00, RMSEA = .02; SRMR = .04; * $p < .05$, ** $p < .01$, *** $p < .001$.

The model reconfirms the result of the analyses of variance reported above in so far as severity of hate speech does not influence participants' intention to counterargue directly ($\beta = -.08$, $p = .15$). Again, it becomes evident that users evaluate hate speech that contains an incitement to violence as a significantly greater threat than hate speech with a lower degree of severity ($\beta = .20$, $p = .003$). However, participants' perception of the hate speech as highly threatening does not increase their intention to intervene ($\beta = -.02$, $p = .77$). Thus, contrary to our assumption, the effect of severity of hate speech on intention to intervene is not mediated by perceived threat ($\beta_{\text{ind\_S1}} = -.004$, $p = .77$), leading us to reject *H1b*.

The more threatening participants assess the hate speech, the more personal responsibility they feel to interfere against it ($\beta = .48$, $p < .001$). Moreover, a higher perceived responsibility leads to an elevated intention to intervene ($\beta = .48$, $p < .001$). Hence, findings reveal that the effect of severity of hate speech on intention to counterargue is mediated by perceived threat and, in turn, by perceived personal responsibility ($\beta_{\text{ind\_S2}} = .05$, $p = .01$). Consequently, *H1c* is supported by our data. The total effect of severity of hate speech on intention to counterargue is not significant ($\beta_{total}$ = –.04, *p* = .47), instead, the total indirect effect ($\beta_{total\_ind}$ = .04, *p* = .04) illustrates the complete mediation of the effect of severity of hate speech by Facebook users' perceptions of threat and of responsibility.

Next, we investigated the effects of the number of bystanders. The results clarify once more that users' intention to counterargue decreases with the number of bystanders being present (β = –.18, *p* = .002), indicating a bystander effect when confronted with hate speech. However, the perception of responsibility remains unaffected by the number of bystanders (β = –.01, *p* = .89). Thus, we found no support for a mediation of the bystander effect by feeling of responsibility ($\beta_{ind\_B}$ = –.004, *p* = .89) as postulated in *H2b*. We therefore have to reject this hypothesis. In conclusion, the total effect of the number of bystanders reached significance ($\beta_{total}$ = –.18, *p* = .003), whereas there was no indirect bystander effect in this study ($\beta_{total\_ind}$ = –.004, *p* = .89).

Consistent with the results reported above, the presence of prior reactions to the hate speech does not influence intention to counterargue (β = .04, *p* = .49). However, users feel a higher personal responsibility to interfere if no one had commented on the hate speech, yet, than if other users had already intervened (β = –.12, *p* = .04). *H3b* posited that the effect of prior reactions to the hate speech on intention to counterargue might be mediated by perceived responsibility, but the indirect effect was only close to significance ($\beta_{ind\_CS}$ = –.06, *p* = .05). Therefore, *H3b* has to be rejected. Hence, the model shows a non-significant total effect of prior reactions ($\beta_{total}$ = –.02, *p* = .78) and a total indirect effect on intention to intervene approaching significance ($\beta_{total\_ind}$ = –.06, *p* = .05). Overall, the model explains a relatively high share of variance of intention to counterargue ($R^2$ = .25, *p* < .001) and of Facebook users' perceived personal responsibility ($R^2$ = .25, *p* < .001) when confronted with hate speech. By contrast, only 4 percent of the variance of perceived threat can be explained by the presumed causal relationships in the model.

## 7. Discussion

For individuals to intervene in a critical situation such as antisocial behavior, they at least have to perceive the situation to be an emergency and personally feel responsible to intervene. However, research shows that there are situational factors hampering the completion of this decision process and leading bystanders to remain passive. In particular, both the degree of severity of the situation and number of bystanders being present have proven to be decisive factors (*bystander effect*, Fischer et al., 2011; Latané & Darley, 1970). Although there is some evidence on key variables shaping bystander intervention in antisocial behavior online, such as cyberbullying or uncivil comments (Naab, Kalch, & Meitz, 2018; Obermaier et al., 2016; Weber, Köhler, & Schnauber-Stockmann, 2018; Weber et al., 2013), up until now only little is known about the factors influencing bystander behavior in hate speech online. However, engaging in counter speech is essential for combatting hate speech and its potential persuasive effects as well as devastating consequences for the groups victimized and for society as a whole, especially when considering the challenging legal situation. Hence, it is highly important to investigate which variables increase the perception of severity as well as the feeling of responsibility to intervene in an incident of hate speech. Our study is trying to provide first insights hereto by focusing both content- and context-related influences on bystander intervention in an incident of hate speech.

Regarding the test of the central steps of the bystander intervention process proposed by Latané and Darley (1970), our study demonstrates that users perceive a highly severe hate speech containing an explicit incitement to violence against asylum seekers to be even more threatening and harmful to the group attacked. However, the perception of severity by itself does not increase their intention to engage in countering. Yet, the more participants deem the hate speech incident to be threatening and, in turn, the more they feel personally responsible to intervene, the more they are willing to counterargue. This association is represented by a weak positive indirect effect. Hence, consistent with evidence on bystander intervention (in cyberbullying), the feeling of responsibility is key to boost the intention to intervene.

Also, results show that a very high compared to a very low number of bystanders does indeed decrease participants' intention to intervene in an incident of hate speech. However, contrary to some of the existing evidence on the bystander effect (in an online environment), the impact of the number of bystanders on intention to counter hate speech is not mediated by the feeling of personal responsibility in our study. There are some possible explanations for this. First, in this study the feeling of responsibility was right on or slightly below the scale midpoint for both a small and a large number of bystanders and, hence, rather low in both conditions. This might be due to the specific characteristics of hate speech: While other forms of antisocial online behavior such as cyberbullying are specifically aimed at an individual herself, hate speech is directed against individuals because of their affiliation to a certain group. That the victim could potentially find support in her group could reduce the individually perceived responsibility of a bystander to take action against hate speech. Hence, the number of bystanders might simply be less powerful in influencing the feeling of responsibility in a hate speech incident than in other forms of antisocial behavior. Second, other variables than the perceived responsibility might be more pronounced in mediating a bystander effect in the context of hate speech. For instance, individuals might be more anxious to engage in countering due to fear of embarrassing themselves or of triggering even more hate the more bystanders have already witnessed a hate speech incident. This fear of evaluation might, in turn, lower their intention to counterargue. Another mediating variable for a bystander effect in an incident of hate speech online might be users' sense of self-efficacy: An individual's believe in her ability to make a change and stop the hater from further spreading hateful messages or change her mind about asylum seekers entirely, again, could boost her intention to counterargue. Therefore, further studies should examine the relation between the number of bystanders and the feeling of responsibility to intervene in hate speech in more detail and, thus, take into account possible alternative mediators of a bystander effect in an incident of hate speech.

Moreover, we demonstrated that the way in which a counter comment has been received by others (endorsed or rejected by a third user), seems to have no implications for feeling of responsibility and intention to counterargue. However, the mere presence of prior reactions of any kind does indeed lower participants' feeling of responsibility and, in turn, their intention to counterargue in comparison to no reaction at all being present. Hence, when seeing other users having already at least uttered one counter comment against the hate speech (no matter if it is retorted), participants seem to detect no need for further interference and deem it unnecessary to spend time and effort of their own to settle the matter. Here again, it could be crucial that hate speech targets a large, diffuse group of victims. This may contribute to bystanders being particularly uncertain whether the hate speech has been read or will ever be read by a member of the victim group at all. Hence, bystanders might be even less likely to intervene when counter speech is already displayed than in other forms of antisocial behavior online. Although testing a rather specific scenario with two comments at the most, this finding is rather contrary to assumptions that expressing counter speech will result in a 'candystorm' by motivating other bystanders to join in countering. Thus, further studies should investigate the conditions under which bystanders do not feel less responsible to intervene with other users countering, but rather motivated to follow suit.

In general, promoting individuals' civic and media literacy could boost their willingness to counter hate online, specifically, by sensitizing citizens to issues relating to hate speech. Civil society initiatives, for instance, might play a pivotal role in teaching online users how to identify hate speech, in what ways counter speech could actually benefit the victimized group (or at the very least avert serious psychological damage), and how to deal with the hater in a constructive manner without having to fear provoking him any further or losing face when trying to intervene in the presence of a large number of bystanders. Current popular examples of such initiatives in the German context are the "#ichbinhier"-movement ("#iamhere" being the English equivalent) on Facebook or a project called "Reconquista Internet" founded by a well-known TV presenter. Both aim at countering hatred in public discourse online by drawing attention to hateful and uncivil content as well as deliberately reacting in an objective and respectful manner to it in order to contribute to an improvement of discussion culture online.

## 8. Limitations and future research directions

When interpreting our results, several limitations must be acknowledged. First, we merely measured users' intention to counterargue by asking if they were willing to comment on the displayed hate speech instead of actually enabling them to write a comment in response and capturing it. This raises the question of external validity with regards to our dependent variable, as it presumably demanded a lot of imagination as well as introspection on behalf of the participants to predict their behavior in a hypothetical situation like this. Therefore, future studies could draw on more externally valid settings, for example, in form of a mixed methods approach that combines content analysis of online users' real reactions to hate speech with individual survey data. That way, participants' purported intention to intervene could be contrasted with their actual behavior in an incident of hate speech online.

Second, we focused on counter speech as a direct online reaction to hate speech. But there are other types of responses against hate speech, which might also be considered counter speech and are available in a real-life incident of hate speech online. These may include confronting the hater face-to-face, participating in demonstrations against xenophobia, or reporting the hater to the platform operators or to the police, and should be accounted for in future studies. Also, the decision to use a single-item capturing the intention to comment against hate speech as dependent variable bears certain disadvantages as it risks reducing the external validity of our measure for countering online and blanks out other forms of counter speech enabled on Facebook (such as sending a private message to the hater(s), adding the emotion ‘angry’ to the hater’s post, or liking a counter speaker’s comment). Nevertheless, we focused on intention to intervene against hate speech *by commenting*, because we consider it the most direct form of public counter speech on SNS. Firstly, countering hate speech in a direct and publicly visible way by commenting assigns specific significance to the number of bystanders witnessing the incident. Hence, in an incident of hate speech on SNS, the influence of the number of bystanders could less discernibly affect the feeling of responsibility or a fear that others evaluate one’s intervention badly if individuals decide to report the hate speech to platform operators (as this would guarantee anonymity) or if they send a private message to the hater (as this would change the setting of the conversation to a private one). Consequently, as our study is one of the first to test the influence of the number of bystanders on the intention to counter a hate speech incident, we chose a single-item measure of countering via commenting in order to avoid varying the influence of number of bystanders across the respondents, contingent on the degree of publicness of their chosen way of intervention. Secondly, only public counter speech can be seen by other bystanders and, in turn, can have a persuasive influence on their opinion, behavior, and perception of public opinion. Hence, countering hate speech by commenting is most valuable for encouraging others to intervene, too, and altering the public discourse. For this reason, we were also encouraged to opt for this most direct form of intervention in hate speech. For different experimental settings, however, we would recommend using a multi-item measurement of counter speech in order to offer a broader range of countering reactions against hate speech to respondents and to represent the complex construct of counter speech more adequately.

Third, generalizations of the absolute strength of the effects detected are certainly limited as we recruited from a convenience sample. Also, an above average proportion of participants were women, academics, and people identifying with the political left—characteristics that contribute to a great homogeneity in the sample and, therefore, limit our results. Hence, it might be worthwhile for future research to replicate our findings and extend them to more heterogeneous samples or specific groups (such as adolescents).

Fourth, our stimulus has specific features. As it represented a hate post on Facebook, countering hateful content on other online sites may be dependent on slightly different factors. Research concerning users' reactions to hate speech should therefore be expanded to other online sites that may provide different visual cues for online users to decide whether to intervene or to ignore the hateful content. Also, we focused on a specific form of social influence respectively the number of bystanders operationalized using the "seen by" tag that Facebook only provides in groups. Hence, research is required to explore the impact of further situational or contextual cues on the central variables in the bystander intervention process (Latané & Darley, 1970), such as the number of "Likes" or the valence of other "Reactions" on Facebook, the degree of anonymity of the users or the personal relationship to the victimized group. Moreover, future research could take into account the influence of further content-based features of hate speech, for instance, the use of expletives or of we-they dichotomy, correctness of grammar, punctuation, and orthography.

Fifth, it must be noted that the rather strict approach to data cleansing resulted in a considerable number of cases being excluded from the analyses.[6] On the one hand, this rather conservative data cleansing can lead to an overestimation of the effects. On the other hand, it is feasible that participants who were excluded due to (completely) inaccurate recalls of the number of bystanders would nevertheless experience effects on the basis of implicit memory (on the role of implicit memory for media effects, e.g., see Yang & Roskos-Ewoldsen, 2007). We argue, however, that in this early phase of research on countering hate speech, it is sensible to focus on explicit memory of the respective content which is why we opted for that strict approach to data cleansing in order to ensure that the effects found are actually due to the individuals' recall of a high or low number of bystanders in the stimulus. However, the effects of implicit memory of content- and context-related influences on bystander intervention in hate speech could be examined in future studies.

Sixth, in addition to the content and contextual factors examined, it is conceivable that factors at the individual level may have an impact on whether individuals

[6] Conducting the analyses with a sample of $n = 444$ in which only participants who did not complete the questionnaire or took too little time to answer it were excluded, we find that the results regarding the main effects remain largely consistent with those reported for the adjusted sample ($n = 304$). First, there was neither a significant main effect on users' willingness to counterargue regarding severity of hate speech ($F(1, 432) = .10$, $p = .75$, $\eta^2_{\text{part}} = .00$; $M_{\text{with incitement to violence}} = 2.11$, $SD = 1.71$, $M_{\text{without incitement to violence}} = 2.06$, $SD = 1.57$), nor prior reactions of other users ($F(2, 432) = .28$, $p = .76$, $\eta^2_{\text{part}} = .001$; $M_{\text{no reaction}} = 2.01$, $SD = 1.73$, $M_{\text{countering reactions}} = 2.09$, $SD = 1.68$, $M_{\text{mixed reactions}} = 2.16$, $SD = 1.67$). A high number of bystanders led to a lower willingness to counterargue ($F(1, 432) = 3.09$, $p = .08$, $\eta^2_{\text{part}} = .01$; $M_{\text{few bystanders}} = 2.23$, $SD = 1.85$, $M_{\text{many bystanders}} = 1.94$, $SD = 1.51$), though the difference was marginally non-significant. Second, analyzing the effects on participants' perception of threat, we found a significant effect of severity of hate speech ($F(1, 432) = 7.37$, $p = .01$, $\eta^2_{\text{part}} = .02$; $M_{\text{with incitement to violence}} = 6.04$, $SD = 1.29$, $M_{\text{without incitement to violence}} = 5.70$, $SD = 1.40$). However, perceived threat was neither affected by number of bystanders ($F(1, 432) = .06$, $p = .80$, $\eta^2_{\text{part}} = .00$; $M_{\text{few bystanders}} = 5.88$, $SD = 1.39$, $M_{\text{many bystanders}} = 5.86$, $SD = 1.34$) nor by prior reactions of other users ($F(2, 432) = .86$, $p = .43$, $\eta^2_{\text{part}} = .004$; $M_{\text{no reaction}} = 5.80$, $SD = 1.42$, $M_{\text{countering reactions}} = 6.00$, $SD = 1.26$, $M_{\text{mixed reactions}} = 5.82$, $SD = 1.39$). Third, concerning feeling of personal responsibility, a significant effect of severity of hate speech emerges ($F(1, 432) = 5.15$, $p = .02$, $\eta^2_{\text{part}} = .01$; $M_{\text{with incitement to violence}} = 4.03$, $SD = 2.06$, $M_{\text{without incitement to violence}} = 3.58$, $SD = 2.01$); yet, number of bystanders ($F(1, 432) = .44$, $p = .51$, $\eta^2_{\text{part}} = .001$; $M_{\text{few bystanders}} = 3.87$, $SD = 2.06$, $M_{\text{many bystanders}} = 3.73$, $SD = 2.04$) as well as prior reactions did not affect it ($F(2, 432) = 1.15$, $p = .32$, $\eta^2_{\text{part}} = .01$; $M_{\text{no reaction}} = 3.99$, $SD = 2.04$, $M_{\text{countering reactions}} = 3.76$, $SD = 2.00$, $M_{\text{mixed reactions}} = 3.65$, $SD = 2.09$).

are willing to counterargue. For example, traits like personality strength (Schenk & Rössler, 2009), self-efficacy (Vecchione & Caprara, 2009), or online engagement (Lin, Chiu, & Luarn, 2015) may affect the willingness to speak up against hateful utterances and should therefore be included in future investigations.

In general, since perceptions of threat and responsibility seem to represent key elements with regard to users' willingness to take action, future research should further address the question what personal or situational factors actually trigger an individual's perception of threat and feeling of personal responsibility when confronted with hate speech and may even explore ways how to enhance these perceptions in the online environment in order to incite countering hate speech. For instance, qualitative interviews or focus groups with online users as well as observing or analyzing the content of interactions in comments beneath hate postings could provide profound insights hereto.

## 9. Conclusion

For the last few years, hate speech on SNS has been a growing concern of platform operators, legislators, and the public in general. While a number of campaigns emerged calling for the individual user to become active against hate online, little research has yet been conducted to identify the external influences of users' willingness to interfere against hateful user-generated content. This study presents key insights into contextual and content-based factors determining Facebook users' willingness to counter hate speech. More specifically, with many other deedless users present, the individual is less likely to intervene herself, indicating the occurrence of a bystander effect in hate speech. Furthermore, the major role of users' perceptions of actual threat for the victimized group of the hate post and of their feeling of personal responsibility is emphasized as prerequisites to engage in countering hate online. Raising awareness to individual users' responsibility when encountering hateful content online could be a first step in order to encourage more counter speech.

## References

Act to Improve Enforcement of the Law in Social Networks (Network Enforcement Act) (2017).

Bagozzi, R., & Yi, Y. (2012). Specification, evaluation, and interpretation of structural equation models. *Journal of the Academy of Marketing Science*, *40*(1), 8–34. https://doi.org/10.1007/s11747-011-0278-x

Bartlett, J., & Krasodomski-Jones, A. (2015). Counter-speech. Examining content that challenges extremism online. Retrieved from http://www.demos.co.uk/wp-content/uploads/2015/10/Counter-speech.pdf

Bastiaensens, S., Vandebosch, H., Poels, K., Van Cleemput, K., DeSmet, A., & De Bourdeaudhuij, I. (2014). Cyberbullying on social network sites. An experimental study into bystanders' behavioural intentions to help the victim or reinforce the bully. *Computers in Human Behavior*, *31*, 259–271. https://doi.org/10.1016/j.chb.2013.10.036

Benesch, S. (2014). *Countering dangerous speech: New ideas for genocide prevention*. Washington, DC: United States Holocaust Memorial Museum. Retrieved from https://www.ushmm.org/m/pdfs/20140212-benesch-countering-dangerous-speech.pdf

Blair, C. A., Foster Thompson, L., & Wuensch, K. L. (2005). Electronic helping behavior: The virtual presence of others makes a difference. *Basic & Applied Social Psychology*, *27*(2), 171–178. https://doi.org/10.1207/s15324834basp2702_8

Boeckmann, R. J., & Liew, J. (2002). Hate speech: Asian American students' justice judgments and psychological responses. *Journal of Social Issues*, *58*(2), 363–381. https://doi.org/10.1111/1540-4560.00265

Costello, M., Hawdon, J., Ratliff, T., & Grantham, T. (2016). Who views online extremism? Individual attributes leading to exposure. *Computers in Human Behavior*, *63*(Supplement C), 311–320. https://doi.org/10.1016/j.chb.2016.05.033

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

Dickter, C. L., & Newton, V. A. (2013). To confront or not to confront: non-targets' evaluations of and responses to racist comments: Responses to racist comments. *Journal of Applied Social Psychology*, *43*, E262–E275. https://doi.org/10.1111/jasp.12022

Ernst, J., Schmitt, J. B., Rieger, D., Beier, A. K., Vorderer, P., Bente, G., & Roth, H.-J. (2017). Hate beneath the counter speech? A qualitative content analysis of user comments on YouTube related to counter speech videos. *Journal for Deradicalization*, (10), 1–49.

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

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

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Kaspar, K., Gräßer, L., & Riffi, A. (Eds.). (2017). *Online Hate Speech: Perspektiven auf eine neue Form des Hasses* [Online hate speech: Perspectives on a new form of hatred]. Düsseldorf München: kopaed.

Latané, B., & Darley, J. M. (1968). Group inhibition of bystander intervention in emergencies. *Journal of Personality and Social Psychology*, *10*(3), 215–221. https://doi.org/10.1037/h0026570

Latané, B., & Darley, J. M. (1970). *The unresponsive bystander: Why doesn't he help?* New York: Appleton-Century-Crofts.

Latané, B., & Nida, S. (1981). Ten years of research on group size and helping. *Psychological Bulletin*, *89*(2), 308–324. https://doi.org/10.1037/0033-2909.89.2.308

Leets, L. (2002). Experiencing hate speech: Perceptions and responses to anti-semitism and antigay speech. *Journal of Social Issues*, *58*(2), 341–361. https://doi.org/10.1111/1540-4560.00264

Leets, L., & Giles, H. (1997). Words as weapons—When do they wound? Investigations of harmful speech. *Human Communication Research*, *24*(2), 260–301. https://doi.org/10.1111/j.1468-2958.1997.tb00415.x

Leiner, D. J. (2016). Our research's breadth lives on convenience samples. A case study of the online respondent pool "SoSci Panel." *Studies in Communication and Media*, *5*(4), 367–396. https://doi.org/10.5771/2192-4007-2016-4-367

Lin, Y.-F., Chiu, Y.-P., & Luarn, P. (2015). Influence of Facebook brand-page posts on online engagement. *Online Information Review*, *39*(4), 505–519. https://doi.org/10.1108/OIR-01-2015-0029

Markey, P. M. (2000). Bystander intervention in computer-mediated communication. *Computers in Human Behavior*, *16*(2), 183–188.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide*. Los Angeles, CA: Author.

Naab, T. K., Kalch, A. & Meitz, T. (2018). Flagging uncivil user comments: Effects of intervention information, type of victim, and response comments on bystander behavior. *New Media & Society*, *20*(2), 777–795. https://doi.org/10.1177/1461444816670923

Nekmat, E., & Gonzenbach, W. J. (2013). Multiple opinion climates in online forums: Role of website source reference and within-forum opinion congruency. *Journalism & Mass Communication Quarterly*, *90*(4), 736–756. https://doi.org/10.1177/1077699013503162

Newman, N., Fletcher, R., Kalogeropoulos, A., Levy, D. A. L., & Nielsen, R. K. (2017). *Reuters Institute Digital News Report 2017*. Oxford, England: Reuters Institute for the Study of Journalism. Retrieved from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/Digital%20News%20Report%202017%20web_0.pdf

Noelle-Neumann, E. (1974). The spiral of silence: A theory of public opinion. *Journal of Communication*, *24*(2), 43–51. https://doi.org/10.1111/j.1460-2466.1974.tb00367.x

Obermaier, M., Fawzi, N., & Koch, T. (2015). Bystanderintervention bei Cybermobbing. Warum spezifische Merkmale computervermittelter Kommunikation prosoziales Eingreifen von Bystandern einerseits hemmen und andererseits fördern [Bystander intervention in cyberbullying. Why characteristics of computer-mediated communication both prevent and promote prosocial intervention of bystanders]. *Studies in Communication and Media*, *4*(1), 28–52. https://doi.org/10.5771/2192-4007-2015-1-28

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

Online Civil Courage Initiative. (n.d.). Retrieved from https://www.facebook.com/Online-CivilCourage/

Palasinski, M. (2012). The roles of monitoring and cyberbystanders in reducing sexual abuse. *Computers in Human Behavior*, *28*(6), 2014–2022. https://doi.org/10.1016/j.chb.2012.05.020

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2016). Measuring the reliability of hate speech annotations: The case of the European refugee crisis. In M. Beißwenger, M. Wojatzki, & T. Zesch (eds.), *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication* (pp. 6–9). Bochum: Bochumer Linguistische Arbeitsberichte.

Schenk, M., & Rössler, P. (2009). The rediscovery of opinion leaders. An application of the personality strength scale. *Communications*, *22*(1), 5–30. https://doi.org/10.1515/comm.1997.22.1.5

Schieb, C., & Preuss, M. (2016). Governing hate speech by means of counterspeech on Facebook. Presented at the 66th Annual Conference of the International Communication Association, Fukuoka, Japan.

Spears, R., & Lea, M. (1992). Social influence and the influence of the 'social' in computer-mediated communication. In M. Lea (Ed.), *Contexts of Computer-Mediated Communication* (pp. 30–65). London: Harvester-Wheatsheaf.

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

Vecchione, M., & Caprara, G. V. (2009). Personality determinants of political participation: The contribution of traits and self-efficacy beliefs. *Personality and Individual Differences*, *46*(4), 487–492. https://doi.org/10.1016/j.paid.2008.11.021

Weber, M., Ziegele, M., & Schnauber, A. (2013). Blaming the victim: The effects of extraversion and information disclosure on guilt attributions in cyberbullying. *Cyber-Psychology, Behavior & Social Networking*, *16*(4), 254–259. https://doi.org/10.1089/cyber.2012.0328

Weber, M., Köhler, C., & Schnauber-Stockmann, A. (2018). Why should I help you? Man up! Bystanders'gender stereotypic perceptions of a cyberbullying incident. *Deviant Behavior*, online first. https://doi.org/10.1080/01639625.2018.1431183

Woong Yun, G., & Park, S.-Y. (2011). Selective posting: Willingness to post a message online. *Journal of Computer-Mediated Communication*, *16*(2), 201–227. https://doi.org/10.1111/j.1083-6101.2010.01533.x

Yang, M., & Roskos-Ewoldsen, D. R. (2007). The effectiveness of brand placements in the movies: Levels of placements, explicit and implicit memory, and brand-choice behavior. *Journal of Communication*, *57*(3), 469–489. https://doi.org/10.1111/j.1460-2466.2007.00353.x

You, L., & Lee, Y.-H. (2018). Bystander effects in cyberbullying on social network sites: Anonymity, group size, and intervention intentions. Paper presented at the 68th Annual Conference of the International Communication Association, Prague, Czech Republic.

Zerback, T., & Fawzi, N. (2017). Can online exemplars trigger a spiral of silence? Examining the effects of exemplar opinions on perceptions of public opinion and speaking out. *New Media & Society*, *19*(7), 1034–1051. https://doi.org/10.1177/1461444815625942