Article

new media & society
1–20

DOI: 10.1177/14614448211017527
journals.sagepub.com/home/nms


I'll be there for you? Effects of Islamophobic online hate speech and counter speech on Muslim in-group bystanders' intention to intervene

Magdalena Obermaier
LMU Munich, Germany

Desirée Schmuck
KU Leuven, Belgium

Muniba Saleem
University of California, Santa Barbara, USA

Abstract

Online hate speech is very common. This is problematic as degrading social groups can traumatize targets, evoke stress, and depression. Since no reaction of others could suggest the acceptability of hate speech, bystander intervention is essential. However, it is unclear when and how minorities react to hate speech. Drawing from social identity theory and research on in-group intervention, we inquire how Islamophobic online hate speech and counter speech by majority or minority members shape Muslims' willingness to intervene. Thus, in an online experiment ($N = 362$), we varied the presence of Islamophobic online hate speech and counter speech by a (non-) Muslim. Results showed that Islamophobic online hate speech led to a perceived religious identity threat which, in turn, increased the personal responsibility to intervene and resulted in higher intentions to utter factual counter speech. In addition, counter speech by both majority and minority members directly reduced Muslims' intentions to counterargue hatefully.

Corresponding author:
Magdalena Obermaier, Department of Media and Communication, LMU Munich, Munich 80538, Germany.
Email: obermaier@ifkw.lmu.de

Keywords
In-group bystander intervention, Islamophobic online hate speech, online bystander intervention, online counter speech, online hate speech, social identity

Religious minorities, and Muslims specifically, are often targeted by hateful online comments (Awan, 2014; Baider et al., 2018). For instance, 40% of German citizens have already witnessed hate speech in digital media, with around 90% of these at least rarely encountering online hate speech against Muslims (Geschke et al., 2019). Hate speech, which disparages others on the basis of "ethnicity, gender, sexual orientation, national origin, or some other characteristic that defines a group" (Hawdon et al., 2017: 254), has been found as being highly problematic. As a potentially traumatizing experience, hate speech may evoke anxiety and stress, lower self-esteem, and trigger depressive thoughts in those affected (Leets, 2002). Frequent confrontation with hate speech can also foster negative stereotypical attitudes of the majority group against minorities (Hsueh et al., 2015), boost polarization tendencies, and aggravate intergroup relations between minority and majority members (Kteily and Bruneau, 2017).

Because no reaction to hateful statements could suggest its acceptability (Kümpel and Rieger, 2019; Sood et al., 2012) and increase the harm for targets (Citron and Norton, 2011), counter speech as a "common, crowd-sourced response" (Bartlett and Krasodomski-Jones, 2015: 5) is essential. It could help targets cope with the incidents (Leets, 2002), prevent them from perceiving that the majority of society agrees with the hate speech (Zerback and Fawzi, 2017), and convince uninvolved users not to endorse the hateful utterances (Schieb and Preuss, 2016). However, to date, it remains largely unclear (a) how the presence of counter speech affects minority members' reactions to online hate speech and (b) which processes determine whether they themselves engage in counter speech.

Therefore, this study advances the literature in three important ways. First, we combine the considerations of social identity theory (Tajfel and Turner, 1986) and the decision model of bystander intervention (Latané and Darley, 1970), to explain the intervention of in-group members (Levine et al., 2002) and, hence, to investigate effects of online hate speech on minorities' counter speech intentions. Specifically, we conceive Islamophobic online hate speech as a threat to Muslims' social identity, which may result in the desire to restore a positive social identity by defending oneself against the hate speech, that is, by uttering counter speech. Second, in doing so, we differentiate two types of online counter speech, *factual* and *hateful* counter speech. Third, drawing from existing research in offline (group-based) bystander intervention, we examine whether the presence of previous counter speech by Muslim in-group members or non-Muslim out-group members accelerates or mitigates these processes. Counter speech might serve as an important indicator for minority members that others do not approve of the hate speech. In particular, counter speech by out-group members may be beneficial for intergroup relations as it could motivate victims to still approach out-group members to enter into constructive discourse (Delgado and Stefancic, 2014).

To that aim, we conducted an online experiment with 362 Muslim participants in Germany to investigate minority intervention in Islamophobic online hate speech and how prior countering reactions of other Muslim minority members or non-Muslim

majority members affect this tendency. Unlike the previous research, which has mainly focused on members of the majority population, our study shifts the perspective by investigating the consequences of online hate speech on those affected by the derogatory content: Muslim social media users.

Hate speech and counter speech

Compared to its growing presence in digital media, research on consequences of online hate speech, especially for targeted groups, is scarce. Hate speech can be defined as statements that "express hatred or degrading attitudes toward a collective" (Hawdon et al., 2017: 254), oftentimes directed against minorities (Keipi et al., 2017). Hence, the main characteristic of hate speech is that, unlike other forms of uncivil online communication (e.g. cyberbullying), it aims to degrade individuals due to their belonging to a particular social, racial, or religious group (Hawdon et al., 2017). *Online* hate speech is distinguished further by features, which are characteristic for the online environment as a potential deindividuation setting, like a perceived sense of anonymity given the difficulty for other users to trace back the origin of the content (Lea et al., 2001; Postmes and Spears, 1998), haters do not get an immediate and full reaction of how hate speech affects minority members and breaks social norms, which could contribute to a lower inhibition threshold for verbal attacks (Suler, 2004). Furthermore, although an act of online hate speech can only mark a single incident (whereas cyberbullying is occurring regularly over a long period of time, Tokunaga, 2010), one incident can lead to repeated victimization due to its oftentimes high reach (Leonhard et al., 2018).

User intervention is considered crucial in combatting online hate (Kümpel and Rieger, 2019). One form of online intervention is *uttering counter speech*, which is defined as "crowd-sourced response to extremism or hateful content" (Bartlett and Krasodomski-Jones, 2015: 5). Although there are many possible ways of countering (Brown, 2016), research on its effects is limited. The few existing findings suggest that while counter speech often may not persuade haters to change their attitudes (Delgado and Stefancic, 2014), it may still shape the discourse norms, opinions, and possibly prosocial behavior of bystanders marking the largest group of witnesses (Leader Maynard and Benesch, 2016; Schieb and Preuss, 2016). Moreover, counterarguing could alleviate the negative impact of hate speech on targets and may motivate targets to still enter into constructive discourse with out-group members (Delgado and Stefancic, 2014; Leets, 2002).

Moreover, literature distinguishes between counter speech that is uttered in a factual or in a hateful way. Examples of *factual counter speech* are labeling hate speech as such, delivering facts and supporting the target group (Bartlett and Krasodomski-Jones, 2015; Naab et al., 2018). Factual counter speech is shown to contribute to a deliberative discussion atmosphere (e.g. mutual respect, openness for different views) increasing the willingness of others to partake, as opposed to humorous or sarcastic responses that mainly provide entertainment value (Ziegele and Jost, 2020). Also, it can convince out-group members of views opposed to the online hate and foster tolerance (Citron and Norton, 2011). In contrast, *hateful counter speech* is defined by degrading the hater. Existing research suggests that hateful counter speech may be counterproductive, as it can cause the hater to more strongly adhere to uncivil attitudes and behaviors (Nyhan and Reifler,

2010). Also, hateful countering could deter others from intervening, and foster an aggressive and even more hostile discourse (Chen and Lu, 2017; Hsueh et al., 2015). Thus, it is important to detect the mechanisms which lead targeted in-group members to intervene against online hate speech in either a factual or hateful way.

Mechanisms of bystander intervention in online hate speech

In order to explain when targets of hate speech intervene themselves, we refer to Latané and Darley's (1970) decision model of bystander intervention, which we will refer to from here on as the bystander intervention model (BIM). For individuals to intervene in antisocial behavior, they must primarily notice a critical situation, assess the situation as threatening (to others and/or self), feel personally responsible for intervening, consider how to help, and (decide to) intervene (Latané and Darley, 1970), whereby some steps can certainly coincide. If any one of these steps is not completed, they are less likely to get involved. The model has already been used to explain bystander intervention in online incivilities in general, cyberbullying, and online hate speech (Leonhard et al., 2018; Obermaier et al., 2016; Weber et al., 2013; Ziegele et al., 2020). Therefore, indications of (parts of) this sequence are also evident for uncivil communication in digital media.

However, the strand of literature on prosocial behavior and the BIM (Latané and Darley, 1970) has found less attention to investigate *in-group* intervention in online hate speech. Although social identification likely contributed to the early findings on bystander intervention (e.g. shared gender identity with Kitty Genovese), the BIM has rarely been integrated with the assumptions of social identity theory (SIT; Turner et al., 1987). Only more recently, a corpus of literature recognized the "importance of exploring the social category relations between all those present in the emergency situation" (Levine et al., 2002: 1453; Dovidio et al., 1997; Levine, 1999) and, thus, incorporated social identity processes in the BIM. In essence, these studies revealed that shared social identities increase bystander intervention but only in certain conditions (Dovidio et al., 2006; Levine and Manning, 2013).

Hence, considering the concept of social identity is crucial when investigating the intervention in online hate speech by minority members. In contrast to online incivilities attacking someone personally (e.g. cyberbullying), hate speech derogates individuals due to characteristics related with a social group (e.g. attacking Malik because he is a Muslim) or a social group as a whole (e.g. attacking Muslims as a group), and, thus, is directed against a social identity (Hawdon et al., 2017). Consequently, by referring to *in-group bystanders*, we focus on "passerby[s]" (Latané and Darley, 1970: 31) that randomly encounter an incident of online hate speech against other in-group members or their in-group as a whole (Levine et al., 2002). Thus, like uninvolved bystanders, in-group members are facing a great deal of uncertainty knowing little about the perpetrator, the situation, and other witnesses. Moreover, in-group bystanders in the offline realm were suspected to be aware that they might share the same fate another time (Dovidio et al., 2006; Levine et al., 2002) and could perceive the incident as an attack on their social group as well. Appropriately, in-group bystanders of online hate speech are (indirectly) targeted by those incidents and, as a result, may feel threatened in their social

identity (Major and O'Brien, 2005). Therefore, we argue that the BIM supplemented by considerations of SIT provides an important theoretical foundation to explain minority intervention and assume that in-group bystanders of online hate speech also have to complete the psychological decision-making process proposed by the BIM to intervene.

Thus, we argue that online hate speech cannot only be understood as an emergency situation that bystanders perceive as more or less threatening (Leonhard et al., 2018), but it may also represent a social identity threat for the targets and for all other social media users who categorize themselves in the addressed in-group (Turner et al., 1987). According to SIT (Tajfel and Turner, 1986), individuals share various social identities, whereby minority members often share various cultural identities, like ethnicity or religion (Statham and Tillie, 2016). These become salient and threatened by situational cues, such as news reports (Saleem and Ramasubramanian, 2019) or targeted political advertising (Schmuck et al., 2017) reflecting one's in-group in a negative light. Hence, those affected by online hate speech assess the threat to the social group to which they feel they belong. Therefore, online hate speech, which is directed against minority groups such as Muslims will likely increase the salience of the threatened cultural identity and motivate coping responses (Major and O'Brien, 2005).

Following the propositions of the BIM (Levine et al., 2002, 2005), for Muslim social media users to intervene in online hate speech against in-group members and given they are aware of a certain incident, they have to perceive the situation as a social identity threat (Major and O'Brien, 2005), must feel personally responsible to intervene, and need to decide (how to) implement their decision. If any one of these steps is not completed, targets are less likely to get involved. Accordingly, experimental studies demonstrate that sharing an in-group identity (Dovidio et al., 1997) or perceiving one's in-group to be threatened (Penner et al., 2005) promotes offline helping toward other in-group members. For instance, individuals identifying with a certain sports team are more likely to help a fellow injured fan than one supporting the opposite team (Levine et al., 2005).

Research on minorities supports that exposure to negative media content can threaten the religious identity of Muslims (Saleem et al., 2020, 2019; Schmuck and Tribastone, 2020; Schmuck et al., 2017). However, Muslim social media users may not necessarily cope with a perceived religious threat evoked by Islamophobic online hate speech by eliciting an action or response toward the hateful content (Leets, 2002). Counterarguing in this context may require that Muslims feel *personally* responsible for intervening (Leonhard et al., 2018). If a response is elicited, it may vary based on whether the content is factual or hateful. This is worthwhile to distinguish as the former can shape a deliberative, the latter a hostile discourse climate (Chen and Lu, 2017; Ziegele and Jost, 2020). Accordingly, we hypothesize,

H1. Compared to no online hate speech, Islamophobic online hate speech will indirectly increase Muslim in-group bystanders' intention to express (a) factual and (b) hateful counter speech, through an increased perception of religious identity threat and, in turn, personal responsibility to intervene.

Influence of online counter speech on in-group intervention in online hate speech

Drawing from research on group-based bystander intervention (Levine et al., 2005, 2010), we are also interested in how the presence of previously uttered online counter speech by

majority members or other minority members affects in-group intervention in online hate speech. Prior experiments suggest that the behavior of other in-group members can shape in-group intervention offline. Thus, individuals are *more likely* to intervene, for instance, in physical violence, if other *in-group* members (e.g. women, students) have already intervened (Levine and Crowther, 2008; Levine et al., 2002). Also, helping or volunteering for in-group purposes increases when imagining oneself in a group of in-group members, but not when bystanders are perceived as out-group members (Levine et al., 2002, 2010).

Regarding counter speech by majority members, for one, there is evidence that out-group bystanders are less likely to intervene in online hate speech when others have already done so, because they may suspect that their help is not needed anymore (Leonhard et al., 2018). Among Muslim minority members, therefore, the presence of out-group counter speech may lead to decreased perceptions that the majority agrees with the online hate speech against minorities (Zerback and Fawzi, 2017) and to an increased appreciation that such statements are not acceptable (Citron and Norton, 2011). Hence, if non-Muslim majority members have already uttered counter speech in response to online hate speech, Muslims may feel less threatened in their religious identity and, in turn, may see less need in expressing counter speech themselves (also see Leets, 2002; Li, 2010). Thus, we hypothesize,

> *H2*. Compared to standalone online hate speech, out-group counter speech in response to Islamophobic online hate speech will indirectly *decrease* Muslim in-group bystanders' intention to express (a) factual and (b) hateful counter speech, through a reduced perception of religious identity threat and, in turn, personal responsibility to intervene.

In contrast, relying on aforementioned research on in-group intervention, with other Muslim minority members intervening in an incident, Muslims could be *more* likely to perceive a strong identity threat and, in turn, should feel more personally responsible to intervene as well (Levine et al., 2005). This is based on findings that individuals rather intend to intervene when (a large number of) other in-group bystanders show(s) a willingness to intervene offline (Levine et al., 2002). This prosocial behavior is attributed to the fact that the presence of other people from the in-group makes prosocial group norms salient and one is more likely to feel the pressure to behave in accordance with the norm (Levine et al., 2010). Thus, we assume,

> *H3*. Compared to standalone online hate speech, in-group counter speech in response to Islamophobic online hate speech will indirectly *increase* Muslim in-group bystanders' intention to express (a) factual and (b) hateful counter speech, through an increased perception of religious identity threat and, in turn, personal responsibility to intervene.

Figure 1(a) and (b) show the full hypothesized models.

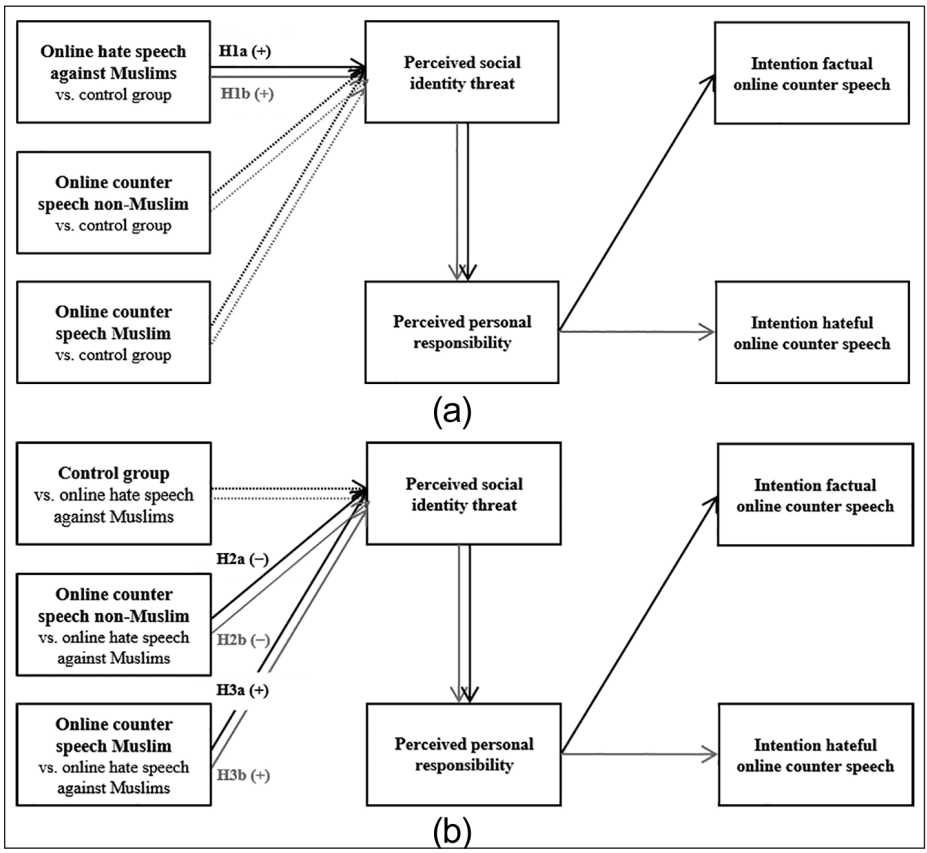

Figure 1. (a) Hypothesized model (control group as reference). (b) Hypothesized model (standalone online hate speech as reference).
Hypothesized indirect effects on the intention to factually counterargue represented in black, on the intention to hatefully counterargue in gray, dotted lines indicate the effect(s) of the remaining independent variables.

Method

Design and participants

We conducted an online experiment with four experimental groups. We varied whether there was (1) standalone Islamophobic online hate speech with no counter speech below, (2) online hate speech with counter speech by a majority member (a non-Muslim) or (3) online hate speech with counter speech by another minority member (a Muslim), and (4) a control group with no online hate speech or counter speech.[1]

Data collection took place in May 2020 and was conducted by a polling company, which recruited panelists resident in Germany who (also) indicated Turkish and/or Arabic as their mother tongue resulting in a total of 426 participants. The total sample

was then adjusted for participants who did not state Islam as their religion, resulting in a sample of $N = 362$ Muslims. Furthermore, we excluded participants with a retention time of one *SD* below the mean dwell time (⩽304 seconds) and those reporting severe difficulties to understand the questionnaire. This resulted in a final sample of $n = 328$ Muslims.[2] All participants consented to the participation and were exposed to a trigger warning at the beginning of the survey, pointing out that it contains potentially distressing material. We also informed them that they may terminate their participation at any point in case they felt uncomfortable. The survey-experiment concluded with a detailed debriefing, which contained elaborate information about the purpose of the study, the nature of the deception (i.e. fictitious posts and comments), and a statement stressing that the hateful statements do not reflect the attitude of the researchers. In addition, the polling company provided participants with contact details for further inquiries about the study. Participation was voluntary and participants received incentives.

Some 48% of the participants were female (average age: 33 years, $SD = 9.80$), 68% had a high school degree, 64% had been living in Germany for more than 10 years and 67% had German as a mother tongue. Also, 41% were born in Germany with at least one parent born abroad, 25% were born abroad themselves; 61% stated to be (very) religious (5-point scale, 1 = "not religious" to 5 = "very religious," agreement to scale points 4 or 5). Furthermore, a randomization check revealed no significant differences between the experimental groups with regard to gender, $\chi^2(3, 328) = 3.66$, $p = .30$, age, $F(3, 324) = 0.60$, $p = .61$, $\eta^2_{part} = .01$; migration background, $\chi^2(6, 328) = 5.73$, $p = .45$; number of years spent in Germany, $F(3, 79) = 0.54$, $p = .66$, $\eta^2_{part} = .02$; prior national identification, $F(3, 324) = 1.28$, $p = .28$, $\eta^2_{part} = .01$; and prior religious identification, $F(3, 324) = 0.54$, $p = .66$, $\eta^2_{part} = .01$.

Stimulus materials

Participants were exposed to a (fictitious) Facebook post by an online outlet of a renowned German quality newspaper promoting an article entitled "Several Daycare Centers eliminate Pork for all Children." The post and the comment(s) were in the Facebook design and due to various features (e.g. mouse over) made as realistically as possible.

The content of the online hate speech and the counter speech was based on previous research (Leonhard et al., 2018; Naab et al., 2018) and the comments were comparable in length (90–100 words).[3] The first comment was written by (fictitious) user "Alex." In the control condition, the comment was neutral stating "Interesting! Muslim kids in daycare," whereas in the treatment conditions it contained Islamophobic hate speech. The hate speech included defamations ("[they] live in their small Muslim parallel world from our taxes") and dehumanization ("these parasites"). Moreover, the hate speech denounced the imposition of the values of the minority group on the majority ("we have to pay for them and then let them tell us how to live") and made a diffuse threat of violence ("we must fight back"; Leader Maynard and Benesch, 2016).

The two counter speech conditions included a second comment either by a majority member ("Mark") or another minority member ("Malik"), whereby their religious affiliation was made clear ("I am/I am not a Muslim"). The comment argued against the hate speech factually ("These are just insinuations against Muslims that are totally

inappropriate and wrong."), flagged it as such, and called upon common values and cohesion ("open society," "we stand together"; Dovidio et al., 2007).

Measures

Participants rated the *degree of severity of the Islamophobic online hate speech* (all dependent variables were measured on 5-point scales, 1 = "do not agree at all" to 5 = "fully agree"), indicating whether it "presents Muslims in a negative light," "presents Muslims as a threat to the German society," "insults Muslims," and "can lead to violence against Muslims" ($\alpha = .89$, $M = 3.83$, $SD = 1.14$).

To measure the *perceived religious identity threat*, participants stated how much they felt "discriminated," "degraded," "offended," and "threatened" by the online hate speech ($\alpha = .92$, $M = 3.35$, $SD = 1.33$; Leonhard et al., 2018; Schmuck and Tribastone, 2020).

Perceived personal responsibility was inquired with the agreement to: "I personally feel obliged to contradict the comment," "I see it as my personal duty to intervene," and "I feel it is my personal responsibility to take action against the commentary" ($\alpha = .87$, $M = 3.27$, $SD = 1.21$; Leonhard et al., 2018; Obermaier et al., 2016).

The *intention to counterargue factually or hatefully* was inquired by asking how participants would react to the online hate speech: "I refute the statements of Alex with factual arguments" and "I write a comment that objectively contradicts the comment of Alex" ($r = .57$***, $M = 3.77$, $SD = 1.04$); "I write a comment that insults non-Muslim Germans" and "I write a comment that offends non-Muslim Germans" ($r = .87$***, $M = 2.39$, $SD = 1.39$).

Results

Treatment check

A treatment check revealed that the degree of severity was estimated to be clearly higher and above the scale midpoint for the Islamophobic online hate speech ($M = 4.11$, $SD = 0.97$) compared to the baseline comment in the control group, which was perceived as fairly neutral ($M = 3.01$, $SD = 1.20$), $t(118.590) = -7.53$, $p < .001$, $d = 1.07$.

Second, a majority correctly recalled whether they saw standalone online hate speech (71%) or a counter comment as well (81%), $\chi^2(1) = 91.87$, $p < .001$. Both groups receiving online counter speech stated that the comment contradicted the hate speech (non-Muslim: $M = 4.09$, $SD = 1.16$; Muslim: $M = 4.01$, $SD = 1.27$), $t(138) = 0.35$, $p = .73$, $d = 0.07$. Interestingly, participants slightly felt that counter speech by a minority member defended Muslims more ($M = 4.30$, $SD = 1.00$) than counter speech by a non-Muslim ($M = 3.91$, $SD = 1.15$), $t(138) = -2.12$, $p = .04$, $d = 0.36$.

Third, a majority correctly indicated that they read online counter speech by a non-Muslim (76%) or Muslim (94%). Thus, the treatment worked adequately.

Data analysis

To test our hypotheses, we applied structural equation modeling using Mplus (Muthén and Muthén, 2010). As independent variables, we dummy-coded the experimental conditions

Table 1. Direct effects of Islamophobic online hate speech and counter speech by majority and minority members (control group as reference).

Predictor	Perceived social identity threat		Perceived personal responsibility		Intention factual online counter speech		Intention hateful online counter speech	
	B	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*
Online hate speech against Muslims[a]	0.91***	0.23	0.12	0.19	–0.06	0.15	0.29	0.31
Online counter speech non-Muslim[a]	1.01***	0.22	0.01	0.17	0.17	0.17	–0.29	0.28
Online counter speech Muslim[a]	1.08***	0.22	0.13	0.16	–0.06	0.17	–0.28	0.28
Perceived social identity threat			0.49***	0.06	–0.01	0.07	–0.17	0.12
Perceived personal responsibility					0.54***	0.12	0.08	0.16
R^2	.10		.34		.47		.05	

$n = 328$, $\chi^2(59) = 84.28$, $p = .02$, $CFI = 0.99$, $RMSEA = 0.04$, $p = .91$, $SRMR = 0.03$.
[a]Control group as reference.
*$p < .05$; **$p < .01$; ***$p < .001$.

into two sets of three variables, where in one set, we used the control group and in the other the standalone hate speech as a reference category. We included the dummy-coded conditions with the control group as reference as manifest variables in the first structural equation model (SEM) and with standalone Islamophobic online hate speech in the second SEM. As dependent latent variables, we used the perceived religious identity threat, the feeling of personal responsibility, and the intention to intervene. Both SEMs offered a very good fit for the data, $\chi^2(59) = 84.28$, $p = .02$, $CFI = 0.99$, $RMSEA = 0.04$, $p = .91$, $SRMR = 0.03$ (Hu and Bentler, 1999). For statistical inference of indirect effects, we used 95% of bias-corrected bootstrap confidence intervals (CI) based on 10,000 samples.

Indirect effects of Islamophobic online hate speech on Muslim in-group bystanders reactions compared to the control group

Hypothesis 1 stated that Islamophobic online hate speech increases Muslims' intention to express counter speech mediated by a perceived religious identity threat and the personal responsibility to intervene. Our first SEM (Table 1, Figure 2) revealed that compared to the control group, standalone Islamophobic online hate speech increased the perceived religious identity threat ($B = 0.91$, $SE = 0.23$, $p < .001$), but it neither affected Muslims' personal responsibility to intervene, nor their intention to counterargue directly. However, the stronger the religious identity threat, the more they felt responsible to intervene ($B = 0.49$, $SE = 0.06$, $p < .001$), which, in turn, increased their intention to counter factually ($B = 0.54$, $SE = 0.12$, $p < .001$), but not hatefully. Accordingly, the results revealed an indirect effect of the standalone online hate speech compared to the control

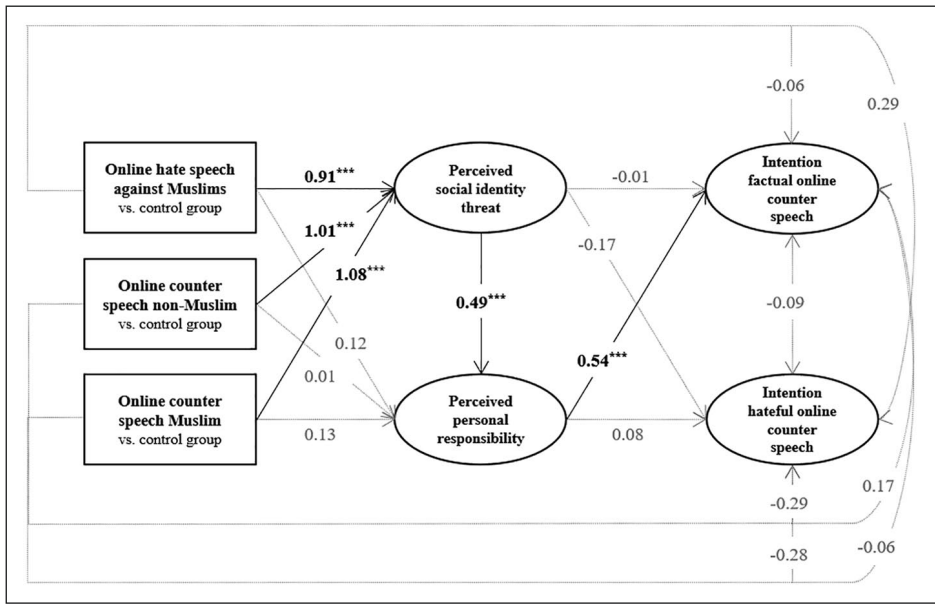

Figure 2. Effects of Islamophobic online hate speech and counter speech by majority and minority members (control group as reference).
$n = 328$, latent variables represented as ovals, manifest variables as rectangles, unstandardized coefficients, significance-testing via bootstrap method (10,000 samples), 95% bias-corrected bootstrap CI, $\chi^2(59) = 84.28$, $p = .02$, $CFI = 0.99$, RMSEA = 0.04, $p = .91$, $SRMR = 0.03$.
$^*p < .05$; $^{**}p < .01$; $^{***}p < .001$.

group on targets' intention to utter *factual counter speech* through increased perceived religious identity threat and, in turn, personal responsibility ($B_{ind} = 0.24$, $SE = 0.08$, CI = [0.11, 0.45]). In addition, Islamophobic online hate speech accompanied by counter speech by both a non-Muslim ($B_{ind} = 0.27$, $SE = 0.09$, CI = [0.13, 0.48]) and another Muslim ($B_{ind} = 0.29$, $SE = 0.09$, CI = [.15, .49]) increased the willingness to counter factually mediated through perceived religious identity threat and personal responsibility.

Thus, irrespective of whether prior counter speech was present or not, Islamophobic online hate speech always led to a religious identity threat, which resulted in higher personal responsibility and higher intentions to counter factually compared to the control group. Hence, H1a was fully supported. However, standalone Islamophobic online hate speech did not indirectly affect targets' intention to *counterargue hatefully*, through a perceived religious identity threat and, in turn, personal responsibility. Therefore, H1b was rejected.[4]

Indirect effects of Islamophobic online hate speech and counter speech on Muslims' in-group bystanders' reactions compared to standalone hate speech

To test whether out-group counter speech in response to Islamophobic online hate speech *decreases* (H2) and in-group counter speech *increases* Muslims' intention to counterargue (H3) mediated by a perceived religious identity threat and personal responsibility, we ran a second SEM using standalone online hate speech as reference (Table 2, Figure 3).

Table 2. Direct effects of control group and online counter speech by majority and minority members (standalone online hate speech as reference).

Predictor	Perceived social identity threat		Perceived personal responsibility		Intention factual online counter speech		Intention hateful online counter speech	
	B	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*
Control group[a]	–0.91***	0.23	–0.12	0.19	0.06	0.15	–0.29	0.31
Online counter speech non-Muslim[a]	0.10	0.21	–0.11	0.18	0.23	0.16	–0.58*	0.27
Online counter speech Muslim[a]	0.17	0.20	0.01	0.18	0.001	0.16	–0.57*	0.28
Perceived social identity threat			0.49***	0.06	–0.01	0.07	–0.17	0.12
Perceived personal responsibility					0.54***	0.12	0.08	0.16
R^2	.10		.34		.47		.05	

$n = 328$, $\chi^2(59) = 84.28$, $p = .02$, $CFI = 0.99$, $RMSEA = 0.04$, $p = .91$, $SRMR = 0.03$.
[a]Standalone online hate speech as reference.
*$p < .05$; **$p < .01$; ***$p < .001$.

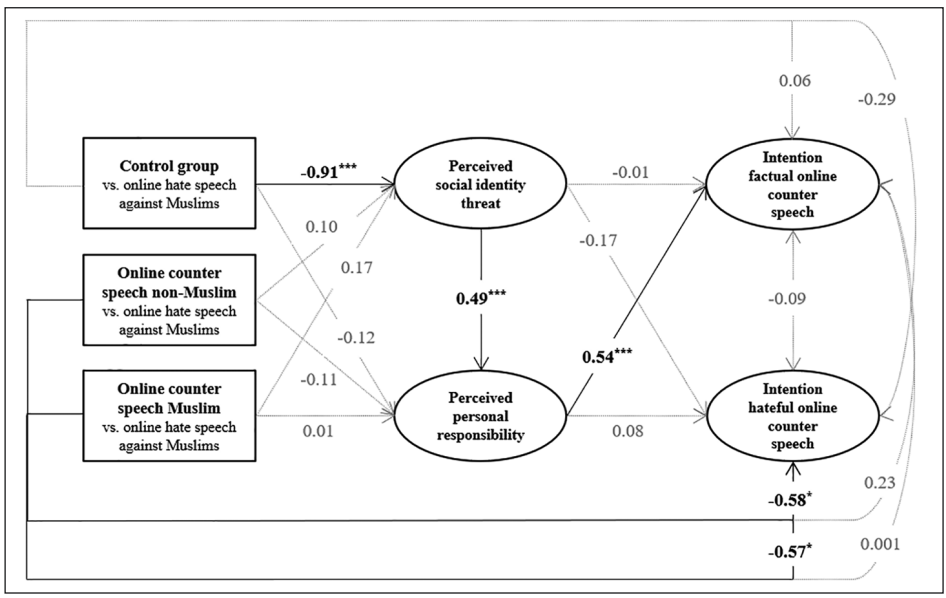

Figure 3. Effects of control group and online counter speech by majority and minority members (standalone online hate speech as reference).
$n = 328$, latent variables represented as ovals, manifest variables as rectangles, unstandardized coefficients, significance-testing through bootstrap method (10,000 samples), 95% bias-corrected bootstrap CI, $\chi^2(59) = 84.28$, $p = .02$, $CFI = 0.99$, $RMSEA = 0.04$, $p = .91$, $SRMR = 0.03$.
*$p < .05$; **$p < .01$; ***$p < .001$.

However, compared to standalone Islamophobic online hate speech, neither majority nor minority counter speech accompanying Islamophobic online hate speech indirectly affected the intention to counterargue factually, mediated by a perceived religious identity threat and, in turn, personal responsibility. Thus, H2a and H3a were rejected. In line with that, there were no indirect effects of counter speech on the willingness to counterargue hatefully. Thus, H2b and H3b were rejected as well.

As already shown in our first SEM, a significant difference between the standalone Islamophobic hate speech compared to the control group emerged ($B = –0.91$, $SE = 0.23$, $p < .001$). Also, replicated from earlier, perceiving a religious identity threat boosted the feeling of personal responsibility and, in turn, the intention to counterargue factually.

Yet, we did find a direct negative effect of online counter speech by either a non-Muslim ($B = –0.58$, $SE = 0.27$, $p = .04$) or another Muslim ($B = –0.57$, $SE = 0.28$, $p = .04$) on targets' willingness to counter hatefully. Thus, the presence of online counter speech reduced Muslims' intention to react in a hateful way directly.

Discussion

The goal of this study was to investigate how Muslim minority members react to Islamophobic online hate speech and whether previously expressed online counter speech by majority or other minority members affects these processes. Taken together, results showed that the confrontation with Islamophobic online hate speech can motivate Muslim in-group bystanders to engage in online counter speech. The underlying mechanisms of this effect are perceived religious identity threat due to the online hate speech, which increases their personal responsibility to intervene as a result. What is also remarkable at this point is that both standalone Islamophobic online hate speech and hate speech accompanied by counter speech was perceived as a threat to their own identity compared to a neutral comment. Thus, prior online counter speech could not reduce the perceived religious identity threat in this study; however, the latter perception mobilized targets to stand up for their social group by engaging in counterarguing themselves. Thus, it remains to be investigated how a larger number of counter-comments affects the level of perceived identity threat by online hate speech for minorities. It is also important to note that perceived responsibility resulted in higher intentions to utter factual, but not hateful online counter speech. In addition, our findings showed that targets' intention to engage in hateful counter speech was in general lower than their intention to use factual counter speech. As a result, when minority members feel responsible to intervene, they decide to do so in a rational, reflected way without degrading the hater.

Contrary to our expectations, we did not detect effects of online counter speech by a majority member or another minority member on the perceived religious identity threat compared to standalone online hate speech. In more detail, when Muslims were confronted with Islamophobic online hate speech that has already been countered by another minority (or majority) member (compared to standalone hate speech), they did not feel more (or less) threatened in their religious identity and, in turn, less personally responsible to intervene. Hence, we could not replicate the findings on group-based bystander intervention offline, that witnessing another in-group member intervene can enhance helping (Levine et al., 2002). There may be several reasons for this; possibly, a singular

counter speech comment from the out-group was not enough support for the in-group bystanders to feel less threatened by the online hate speech compared to standalone hate speech. Similarly, it could be that one single counter comment from the in-group could not yet increase the perceived threat of online hate speech against their own group. It is therefore conceivable that a larger number of counter speech comments could bring differences to light here. These findings could also be due to the nature of the online context. Whereas a larger number of in-group bystanders could more easily defend a physically attacked fellow, in-group online bystanders might perceive that their intervention is no longer needed when counter speech has already been written. Yet, they would then be more likely to intervene in standalone online hate speech (Leonhard et al., 2018), which did not emerge. Thus, because it may take several counter comments to sufficiently suggest that intervention is necessary (Delgado and Stefancic, 2014) and to lead to those affected feeling supported and encouraged to intervene themselves, follow-up studies should vary a higher number of counter speech comments and explicitly survey the perceived efficacy of and the perceived support from the intervention.

Moreover, Levine et al. (2010) suggest that in-group intervention is boosted with an increasing number of in-group bystanders. In addition, they found that it also depends on the identity-specific norms regarding prosocial behavior whether a salient in-group identity boosts helping. However, since our study is one of the first on in-group intervention in online hate speech against minorities, we did not inquire Muslims' salient in-group norms regarding counterarguing to support their group. Thus, follow-up studies should vary the number of minority (and majority) online bystanders and include salient in-group norms regarding group-based support to investigate these mechanisms in more detail. Here, it was only the presence of standalone online hate speech (compared to a neutral comment) that indirectly increased their intention to constructively intervene.

We did find, however, that the presence of online counter speech directly reduced Muslim in-group bystanders' intentions to engage in hateful counter speech, regardless of whether the counter speech has been uttered by a majority member or another minority member. Although online hate speech accompanied by counter speech still exerts a threat to Muslims' social identity as evident in our results, it may reduce the aggressive cognitions, emotions, and behavioral intentions to defend their religious group (Hsueh et al., 2015). This finding is especially relevant as research suggests that hateful counter speech can provoke haters to engage in even more uncivil behavior (Chen and Lu, 2017). Consequently, hateful online counter speech may result in a reinforcing spiral of hostility harming intergroup relations lastingly. Hence, the notion that online counter speech reduces Muslims' tendencies to counter hatefully is a key finding of our study, which has important practical implications. However, the mechanisms affecting the intention to counter hate other than by the variables proposed by the BIM have yet to be discovered in more detail.

Overall, this study is the first to investigate the attitudinal outcomes and behavioral intentions of Muslim minority members in response to Islamophobic online hate speech. More specifically, the study utilizes the BIM in order to explain in-group intervention (Levine et al., 2002) for the first time in online hate speech against minorities. In line with research on bystander intervention (Leonhard et al., 2018; Naab et al., 2018; Obermaier et al., 2016; Ziegele et al., 2020), we show that personal responsibility as an

underlying mechanism also plays a central role when considering intervention in online hate speech by a victimized minority. Although we find that hate speech (accompanied by counter speech) represents a social identity threat for Muslims, they seem to canalize this threat in a constructive way by reacting with factual counter speech. While social identity threats can have adverse effects on the integration of Muslim immigrants (Saleem et al., 2019; Schmuck et al., 2017), it can also motivate collective action to improve the in-group's status in society (Saleem et al., 2020). Similarly, our findings revealed that when confronted with hate, Muslim users are willing to respond in a factual, but not hateful, online counter speech which is important for sustaining positive intergroup relations in a multicultural democracy.

Limitations and future research

Of course, this study has some notable limitations. First, to explain in-group bystander intervention of the Muslim minority in online hate speech, we referred to the BIM and related studies on in-group intervention offline (Levine and Manning, 2013). In addition to leaving out the step of recognizing the emergency due to the experimental setting, we did not assess participants' actual comments or thoughts of how best to help in this context. Instead, we assessed intentions to counter-argue factually or hatefully which is a proxy to both type of assistance and actual behavioral assistance. Future studies can examine the role of emphasizing the benefits of counter speech and high levels of political and writing self-efficacy which could foster greater intervention (Ziegele et al., 2020).

Second, we have not distinguished between degrees of severity of the online hate speech. Since previous findings show that especially the perceived severity of an incident indirectly increases online bystander intervention (Leonhard et al., 2018; Wang, 2020), future research should investigate whether the degree of severity of online hate speech also determines targets' intervention. Also, we did not manipulate the number of likes, "seen by" or shares of our posts, which could influence perceived responsibility to help (Obermaier et al., 2016). Thus, follow-up studies should vary indicators of how many other users have already witnessed the incident and investigate the effects of majority and minority counter speech against this backdrop (Levine et al., 2010).

Third, our results need to be interpreted against the background of the type of counter speech we used, which appealed to pluralistic and tolerant values in democracies. Thus, follow-up studies should vary different types of counter speech (e.g. factual, hateful, or humorous, Ziegele and Jost, 2020) by majority and minority members and investigate its outcomes on target intervention and intergroup relations.

Fourth, since our goal was to investigate how online hate speech affects minority members' attitudinal outcomes and behavioral intentions, we only varied the religious identity of the counter speaker, but not that of the other bystanders. Therefore, future research should examine whether the intention to counterargue varies as a function of the congruence of targets' and bystanders' religious identity (Levine and Crowther, 2008). Finally, we did not measure participants' prior exposure to online hate speech or their previous engagement in counter speech, because surveying this immediately before the manipulation could lead to different responses biasing the results. Thus, follow-up

studies should take prior experiences with online hate and counter speech into account and investigate how they affect the results of this study. This could be done with a multiple-exposure experiment (Schmuck and Tribastone, 2020), allowing for an assessment of prior experiences and responses to the stimuli at different time points.

Implications

Despite these limitations, our findings have important practical implications. For one, given that the perceived responsibility to intervene in online hate speech can foster factual counter speech, as a society we should equip citizens and especially adolescents with the skills to recognize hate speech and to know about means to do something about it. Our finding that factual online counter speech can prevent targets from hating back and, thus, prevent intergroup distancing in a larger scale underlines the relevance for media literacy and prevention campaigns that educate about the characteristics of hate speech against minorities and its consequences for the targeted groups, as well as on how to provide support to facilitate coping.

For another, our findings tackle the issue whether in-group online intervention can be worthwhile. In-group intervention in online hate speech could backfire, provoking even more hostile utterances and leading targets to suffer from a loss of self-efficacy in a conflict that seemingly cannot be solved (Delgado and Stefancic, 2014). Yet, intervention could not only prevent hateful counter speech by minority members, but increase the motivation for countering negative rhetoric against marginalized groups from many different perspectives (Lane et al., 2019; Roden & Saleem, 2021). In other words, minority and majority intervention in online hate speech could promote the importance of being there for each other.

Acknowledgements

The authors would like to thank Julian Unkel for allowing them to use his unpublished, interactive Facebook mockup for Sosci Survey (Leiner, 2019). They are also very grateful for his support in customizing the programming for their project.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Magdalena Obermaier https://orcid.org/0000-0002-3055-3744

Desirée Schmuck https://orcid.org/0000-0002-9492-6052

Supplemental material

Supplemental material for this article is available online.

Notes

1. The dataset with the variables we used in this study is available at: https://osf.io/fmauv/
2. According to a power analysis ($1-\beta = .80$, $p < .05$), the sample size is sufficient to detect moderate effects, which is adequate to trace the direct effects of the variables of interest suggested in the bystander intervention model (Leonhard et al., 2018; Obermaier et al., 2016; Ziegele et al., 2020).
3. For an overview of the stimulus materials, see Table 3 (Supplemental material).
4. For all indirect effects on the intention to counterargue, see Tables 4 and 5 (Supplemental material).

References

Awan I (2014) Islamophobia and Twitter. A typology of online hate against Muslims on social media. *Policy & Internet* 6: 133–150.

Baider FH, Constantinou A and Petrou A (2018) The conceptual contiguity of race and religion. In: Assimakopoulos S, Baider FH and Millar S (eds) *Online Hate Speech in the European Union. A Discourse-Analytic Perspective*. Berlin: Springer, pp. 72–77.

Bartlett J and Krasodomski-Jones A (2015) Counter-Speech. Examining content that challenges extremism online. Available at: http://www.demos.co.uk/project/counter-speech/ (accessed 21 April 2021).

Brown R (2016) Defusing hate: a strategic communication guide to counteract dangerous speech. *US Holocaust Memorial Museum*. Available at: https://www.ushmm.org/m/pdfs/20160229-Defusing-Hate-Guide.pdf (accessed 21 April 2021).

Chen GM and Lu S (2017) Online political discourse: exploring differences in effects of civil and uncivil disagreement in news website comments. *Journal of Broadcasting & Electronic Media* 61(1): 108–125.

Citron DK and Norton HL (2011) Intermediaries and hate speech: fostering digital citizenship for our information age. *Boston University Law Review* 91: 1435–1484.

Delgado R and Stefancic J (2014) Hate speech in cyberspace. *Wake Forest Literature Review* 49: 319–343.

Dovidio JF, Gaertner SL and Saguy T (2007) Another view of "we": majority and minority group perspectives on a common in-group identity. *European Review of Social Psychology* 18: 296–330.

Dovidio JF, Gaertner SL, Validizic A, et al. (1997) Extending the benefit of recategorization: evaluations, self-disclosure, and helping. *Journal of Experimental Social Psychology* 33: 401–442.

Dovidio JF, Piliavin JA, Schroeder DA, et al. (2006) *The Social Psychology of Prosocial Behavior*. Mahwah, NJ: Lawrence Erlbaum.

Geschke D, Klaßen A, Quent M, et al. (2019) #Hass im Netz: Der schleichende Angriff auf unsere Demokratie. Eine bundesweite repräsentative Untersuchung. Available at: https://blog.campact.de/wp-content/uploads/2019/07/Hass_im_Netz-Der-schleichende-Angriff.pdf (accessed 21 April 2021).

Hawdon J, Oksanen A and Räsänen P (2017) Exposure to online hate in four nations. A cross-national consideration. *Deviant Behavior* 38: 254–266.

Hsueh M, Yogeeswaran K and Malinen S (2015) "Leave your comment below." Can biased online comments influence our own prejudicial attitudes and behaviors? *Human Communication Research* 41: 557–576.

Hu L and Bentler PM (1999) Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal* 6: 1–55.

Keipi T, Näsi MJ, Oksanen A, et al. (2017) *Online Hate and Harmful Content: Cross-National Perspectives*. New York, NY: Routledge.

Kteily N and Bruneau E (2017) Backlash: the politics and real-world consequences of minority group dehumanization. *Personality and Social Psychology Bulletin* 43: 87–104.

Kümpel A and Rieger D (2019) *Wandel Der Sprach- Und Debattenkultur in Sozialen Online-Medien*. Available at: https://www.kas.de/de/einzeltitel/-/content/wandel-der-sprach-und-debattenkultur-in-sozialen-online-medien (accessed 21 April 2021).

Lane DS, Coles SM and Saleem M (2019) Solidarity effects in social movement messaging: how cueing dominant group identity can increase movement support. *Human Communication Research* 45: 1–26.

Latané B and Darley JM (1970) *The Unresponsive Bystander: Why Doesn't He Help?* New York, NY: Appleton-Century-Crofts.

Lea M, Spears R and de Groot D (2001) Knowing me, knowing you: anonymity effects on social identity processes within groups. *Personality and Social Psychology Bulletin* 27: 526–537.

Leader Maynard J and Benesch S (2016) Dangerous speech and dangerous ideology: an integrated model for monitoring and prevention. *Genocide Studies and Prevention: An International Journal* 9: 70–95.

Leets L (2002) Experiencing hate speech. Perceptions and responses to anti-Semitism and antigay speech. *Journal of Social Issues* 58: 341–361.

Leiner DJ (2019) *SoSci Survey* (Version 3.1.06) [Computer software]. Available at: https://www.soscisurvey.de

Leonhard L, Rueß C, Obermaier M, et al. (2018) Perceiving threat and feeling responsible. How severity of hate speech, number of bystanders, and prior reactions of others affect bystanders' intention to counterargue against hate speech on Facebook. *Studies in Communication and Media* 7: 555–579.

Levine M (1999) Rethinking bystander non-intervention: social categorization and the evidence of witnesses at the James Bulger murder trial. *Human Relations* 52: 1133–1155.

Levine M and Crowther S (2008) The responsive bystander: how social group membership and group size can encourage as well as inhibit bystander intervention. *Journal of Personality and Social Psychology* 95: 1429–1439.

Levine M and Manning R (2013) Social identity, group processes, and helping in emergencies. *European Review of Social Psychology* 24(1): 225–251.

Levine M, Cassidy C and Jentzsch I (2010) The implicit identity effect: identity primes, group size, and helping. *British Journal of Social Psychology* 49: 785–802.

Levine M, Cassidy C, Brazier G, et al. (2002) Self-categorization and bystander non-intervention. Two experimental studies. *Journal of Applied Social Psychology* 32: 1452–1463.

Levine M, Prosser A, Evans D, et al. (2005) Identity and emergency intervention: how social group membership and inclusiveness of group boundaries shapes helping behavior. *Personality and Social Psychology Bulletin* 31: 443–453.

Li Q (2010) Cyberbullying in high schools. A study of students' behaviors and beliefs about this new phenomenon. *Journal of Aggression, Maltreatment & Trauma* 19: 372–392.

Major B and O'Brien LT (2005) The social psychology of stigma. *Annual Review of Psychology* 56: 393–421.

Muthén LK and Muthén BO (2010) *Mplus User's Guide*. Los Angeles, CA: Muthén & Muthén.

Naab TK, Kalch A and Meitz T (2018) Flagging uncivil user comments: effects of inter-vention information, type of victim, and response comments on bystander behavior. *New Media & Society* 20(2): 777–795.

Nyhan B and Reifler J (2010) When corrections fail. The persistence of political misperceptions. *Political Behavior* 32: 303–330.

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

Penner LA, Dovidio JF, Piliavin JA, et al. (2005) Prosocial behavior: multilevel perspectives. *Annual Review of Psychology* 56: 365–392.

Polder-Verkiel SE (2012) Online responsibility: bad Samaritanism and the influence of internet mediation. *Science and Engineering Ethics* 18: 117–141.

Postmes T and Spears R (1998) Deindividuation and anti-normative behavior: a meta-analysis. *Psychological Bulletin* 123: 238–259.

Roden J and Saleem M (2021) White apathy and allyship in uncivil racial social media comments. *Mass Communication and Society*. Epub ahead of print 6 August. DOI: 10.1080/15205436.2021.1955933

Saleem M and Ramasubramanian S (2019) Muslim Americans' responses to social identity threats: effects of media representations and experiences of discrimination. *Media Psychology* 22(3): 373–393.

Saleem M, Hawkins I, Wojcieszak M, et al. (2020) When and how negative media representations empower collective action in minorities. *Communication Research* 48: 291–316.

Saleem M, Wojcieszak ME, Hawkins I, et al. (2019) Social identity threats: how media and discrimination affect Muslim Americans' identification as Americans and trust in the US government. *Journal of Communication* 69(2): 214–236.

Schieb C and Preuss M (2016) Governing hate speech by means of counterspeech on Facebook. In: *6th Annual Conference of the International Communication Association*, Fukuoka, Japan, 9–13 June.

Schmuck D and Tribastone M (2020) Muslims take action. How exposure to anti-Islamic populist political messages affects Young Muslims' support for collective action: a longitudinal experiment. *Political Communication* 37: 635–655.

Schmuck D, Matthes J and Paul FH (2017) Negative stereotypical portrayals of Muslims in right-wing populist campaigns. Perceived discrimination, social identity threats, and hostility among young Muslim adults. *Journal of Communication* 67: 610–634.

Sood S, Antin J and Churchill E (2012) Profanity use in online communities. In: *Proceedings of the SIGCHI Conference on human factors in computing systems*, Austin, TX, 5–10 May, pp. 1481–1490. New York, NY: ACM Press.

Statham P and Tillie J (2016) Muslims in their European societies of settlement: a comparative agenda for empirical research on socio-cultural integration across countries and groups. *Journal of Ethnic and Migration Studies* 42(2): 177–196.

Suler J (2004) The online disinhibition effect. *Cyberpsychology & Behavior* 7: 321–326.

Tajfel H and Turner JC (1986) The social identity theory of intergroup behavior. In: Worchel S and Austin WG (eds) *Psychology of Intergroup Relations*. Chicago, IL: Nelson-Hall, pp. 7–24.

Tokunaga RS (2010) Following you home from school. A critical review and synthesis of research on cyberbullying victimization. *Computers in Human Behavior* 26: 277–287.

Turner JC, Hogg MA, Oakes PJ, et al. (1987) *Rediscovering the Social Group: A Self-Categorization Theory*. Oxford: Blackwell.

Wang S (2020) Standing up or standing by: bystander intervention in cyberbullying on social media. *New Media & Society* 23: 1379–1397.

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

Zerback T and 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: 1034–1051.

Ziegele M and Jost PB (2020) Not funny? The effects of factual versus sarcastic journalistic responses to uncivil user comments. *Communication Research* 47(6): 891–920.
Ziegele M, Naab TK and Jost P (2020) Lonely together? Identifying the determinants of collective corrective action against uncivil comments. *New Media & Society* 22: 731–751.

Author biographies

Magdalena Obermaier is a research associate at the Department of Media and Communication at LMU Munich, Germany. Her research focuses on media effects, specifically effects of uncivil and prosocial communication in digital media, effects of persuasive communication, and the relationship between journalism and public relations.

Desirée Schmuck is an assistant professor at Leuven School for Mass Communication Research at KU Leuven, Belgium. Her research investigates digital media effects on individuals' well-being in different areas like political, mobile, or environmental communication.

Muniba Saleem is an associate professor at the Department of Communication at University of California, Santa Barbara. Her research examines how media depictions of minorities influence intergroup relations and identity dynamics.