Article

European Journal *of* Criminology

European Journal of Criminology
2023, Vol. 20(3) 817–833

DOI: 10.1177/14773708231156328
journals.sagepub.com/home/euc
Sage

Too civil to care? How online hate speech against different social groups affects bystander intervention

Magdalena Obermaier, Ursula Kristin Schmid, and Diana Rieger
LMU Munich, Germany

Abstract

A large share of online users has already witnessed online hate speech. Because targets tend to interpret such bystanders' lack of reaction as agreement with the hate speech, bystander intervention in online hate speech is crucial as it can help alleviate negative consequences. Despite evidence regarding online bystander intervention, however, whether bystanders evaluate online hate speech targeting different social groups as equally uncivil and, thereby, equally worthy of intervention remains largely unclear. Thus, we conducted an online experiment systematically varying the type of online hate speech as homophobia, racism, and misogyny. The results demonstrate that, although all three forms were perceived as uncivil, homophobic hate speech was perceived to be less uncivil than hate speech against women. Consequently, misogynist hate speech, compared to homophobic hate speech, increased feelings of personal responsibility and, in turn, boosted willingness to confront.

Keywords

Bystander intervention, homophobia, incivility, misogyny, online hate speech, racism

About 80% of German internet users have already witnessed hate speech online (Isenberg, 2019; LfM, 2022). Such speech typically derogates others based on their 'ethnicity, gender, sexual orientation, national origin, or some other characteristic that defines a group' (Hawdon et al., 2017: 254). Because individuals tend to interpret a lack of reaction by others to indicate that they agree with the hateful statements, uninvolved

Corresponding author:
Magdalena Obermaier, Department of Media and Communication, LMU Munich, Oettingenstr. 67, 80538 Munich, Germany.
Email: obermaier@ifkw.lmu.de

bystanders, as the largest group of witnesses of these incidents, play a central role (Leonhard et al., 2018; Schieb and Preuss, 2016). For the targeted social groups and minorities, the absence of interventions can have severe consequences. For instance, there is evidence that hate speech—similar to traumatizing events—can trigger negative emotions and physiological reactions (for example, fear, stress) and foster negative cognitions, such as depressive thoughts and a loss of self-esteem (Geschke et al., 2019; Leets, 2002). In addition, we know that witnessing hate crimes offline can exacerbate one's negative attitudes toward targeted groups (Keel et al., 2022). Thus, bystander intervention in online hate speech is critical because it is thought to help alleviate negative consequences for targeted groups (Preuß et al., 2017). This holds true especially for counter-arguing in publicly visible user comments (Bartlett and Krasodomski-Jones, 2015), as this particular form of intervention may prevent both targets and other bystanders from perceiving that online hate speech is acceptable in a democratic society (Kümpel and Rieger, 2019; see also Zerback and Fawzi, 2017). Thereby, it may also help to prevent hateful online discourses (Hsueh et al., 2015) and polarization tendencies (Meleagrou-Hitchens and Kaderbhai, 2017) in the long run. In this vein, it is argued that there is a moral imperative for bystander intervention in online hate speech because it openly rejects and counteracts verbal attacks using inclusionary messages about the targeted groups (Iganski, 2020).

Some preliminary evidence suggests that uninvolved bystanders intervene to dispute online hate speech the more threatening they believe it to be (Leonhard et al., 2018). However, whether bystanders evaluate online hate speech against different social groups as equally uncivil or threatening and, thus, equally worthy of intervention remains largely unclear to date. Our study addresses this question and, thus, extends the existing literature on several points. To the best of our knowledge, we are among the first to systematically compare the perceived incivility of hate speech against different social groups. To do so, we focus on three of the groups most frequently affected by online hate speech (Geschke et al., 2019; Nennstiel and Isenberg, 2022): individuals with a migration background (that is, people who themselves and/or at least one of their parents and/or grandparents were not born in their country of residence), women, and the LGBTQIA+ community. Second, we investigate whether witnessing one of these types of hate speech can increase the intention to provide messages countering such speech, and, third, we examine the mechanisms behind this intervention referring to the decision model of bystander intervention (here referred to as the bystander intervention model or BIM; see Latané and Darley, 1970). For this purpose, we conducted an online experiment whereby we varied the type of online hate speech as homophobia, racism, or misogyny in order to study how online hate speech against different social groups can affect the intention of uninvolved bystanders to counter such speech.

Online hate speech and counterspeech

As the smallest common denominator states that 'express hatred or degrading attitudes toward a collective' (Hawdon et al., 2017: 254) are defined as hate speech. Thereby, hate speech devalues individuals on the basis of personal characteristics by which they can be assigned to certain social groups such as race, gender, and sexual orientation

(Hawdon et al., 2017; Keipi et al., 2017; Schwertberger and Rieger, 2021). This differs from other forms of uncivil communication in digital media that exceed social norms of interpersonal communication (Kümpel and Rieger, 2019; Ziegele et al., 2020) or that personally attack individuals (as in cyberbullying) without necessarily disparaging their social group memberships (Wachs et al., 2021). Hence, online hate speech derogates and threatens individuals on the basis of their social identity (Major and O'Brien, 2005; Tajfel and Turner, 1986) and is often—but not exclusively—directed against social minorities, as we will show. Moreover, even one incident of hate speech can result in repeated victimization of targets as (public) utterances often have extensive reach in digital media, whereas for cyberbullying, long-term exposure is necessary to give it that label (Tokunaga, 2010). This clearly distinguishes online hate speech from cyberbullying, although both are forms of uncivil online communication and, in reality, can occur concurrently (Schwertberger and Rieger, 2021; Wachs et al., 2021).

Counterspeech as a 'crowd-sourced response to extremism or hateful content' (Bartlett and Krasodomski-Jones, 2015: 5) is often considered an essential form of intervention in hate speech. This is because counterspeech by uninvolved bystanders is thought to help mitigate negative consequences for targets. Although it may not succeed in persuading or silencing haters, it still has the potential to shape discourse norms and to show other uninvolved witnesses that hate speech does not correspond to the majority opinion and encourage them to intervene as well (Bartlett and Krasodomski-Jones, 2015). Furthermore, it could lead to the counterarguing of different perspectives of greater scope and range, which is presumed to be highly effective in combating hate speech (Delgado and Stefancic, 2014).

Mechanisms of bystander intervention in online hate speech

We utilize the bystander intervention model (BIM; Latané and Darley, 1970) to explain bystander intervention in online hate speech against different social groups. The model was originally developed to explain why bystanders do not intervene in face-to-face emergency situations. According to the BIM, uninvolved bystanders must, primarily, (1) assess a situation as threatening and, thus, an emergency for those affected; (2) perceive that they are personally responsible for intervening; and finally, (3) make the decision to intervene and act accordingly (Latané and Darley, 1970). According to this model, bystanders will be less inclined to intervene whether one of these steps is not (sufficiently) taken.

This step-by-step process has previously been described to understand bystander intervention in different forms of uncivil online communication (Bastiaensens et al., 2014; Obermaier et al., 2016; Ziegele et al., 2020) and more specifically in relation to online hate speech (Leonhard et al., 2018). Thus, it has been shown that bystanders of cyberbullying, uncivil user comments, and hate speech do not necessarily intervene more often when they assess an incident as more threatening (see, however, Koehler and Weber, 2018). Rather, it appears that they must, in turn, feel a personal responsibility to intervene in order for them to decide, ultimately, to do so (Leonhard et al., 2018; Obermaier et al., 2016; Ziegele et al., 2020). In line with that, a more severe incident of hate speech (for example, with a threat of violence) has been demonstrated to enhance bystanders' feelings of personal responsibility and, thereby, increase their willingness to intervene with counterspeech (Leonhard et al., 2018).

To date, research on prosocial bystander intervention in uncivil comments in digital media, in general, and hate speech, in particular, has not systematically considered different characteristics that promote an assessment of hate speech as uncivil. This could affect feelings of personal responsibility to intervene to help the targeted group and, ultimately, increase the likelihood of intervention. Therefore, the focus of the present study is to investigate whether not only certain features of the comments lead bystanders to assess them as a threat to those affected (for example, violence, name-calling) but also whether it is the reference to a particular social group that prompts bystanders to intervene.

Bystander intervention in online hate speech against different social groups

As far as bystanders' perceptions of online hate speech against different social groups is concerned, several studies of online bystander intervention have addressed the linguistic or content features that may influence the perceived threat or incivility of online hate speech. For instance, there is some evidence that online hate speech that includes a threat of violence is perceived to be more harmful to targets and/or intervening bystanders than hate speech without such a threat (Leonhard et al., 2018). Moreover, according to a study of people who collectively counterargue against hate speech in social media, user comments containing insults, threats, and stereotypes are perceived to be the most harmful (Ziegele et al., 2020). Similarly, several studies on uncivil utterances (online) in general have suggested that bystanders deem statements containing insulting language, name-calling (Kenski et al., 2020), threats of violence, and insults as the most harmful and uncivil (Stryker et al., 2016).

Studies considering targets' perceptions of statements of discrimination or hate speech (online or off) suggest that hatred directed against different social groups can be perceived as variously threatening, discriminatory, or uncivil. However, most studies have focused on a single targeted group at a time (McCoy and Major, 2003; Sellers et al., 2003). An exception, Leets (2002) demonstrated that homophobic offline hate speech leads to more negative emotional responses by those targeted (for example, fear, anger), whereas anti-Semitic speech results in more behavioral reactions (for example, hating back). Relatedly, women as targets are also more affected by the respective discriminatory hate speech than men (Wojatzki et al., 2018).

In Germany, most commonly, online hate speech relates to targets' sexual orientation (for example, homophobia), ethnicity (that is, racism), and female gender (that is, misogyny) (Döring and Mohseni, 2020; Geschke et al., 2019). Regarding misogynist online hate speech, it is again evident that women are more frequently affected by such utterances than men (Döring and Mohseni, 2019; Mohseni, 2021). These forms are also frequently studied social groups in research on the effects of online hate speech on bystanders' evaluations of the targeted group (Fasoli et al., 2016; also see Nielsen, 2002) and online bystander intervention (Leonhard et al., 2018;

Lück and Nardi, 2019; Wilhelm et al., 2020; Wilhelm and Joeckel, 2019; Ziegele et al., 2018).

Therefore, in the present study, we aim to compare bystanders' perceptions of and interventions in online hate speech in relation to these three social groups (LGBTQIA+, individuals with a migration background, and women). A major difference between hate speech directed toward these groups is that women are usually not described as a 'threatened' social group because they are not considered a social minority (for a critical overview, see Steinl, 2019). Consequently, although women are among the primary targets of online hate speech, German law does not provide for the prosecution of misogynist hate speech. Although numerous different forms of gender-based online incivility toward women are found and discussed (Henry and Powell, 2018), gender-based and especially misogynist hate speech maintains a special role—not only in Germany but also in regard to international human rights law (Sękowska-Kozłowska et al., 2022). Debating whether or not hatred toward women can be classified as hate speech (for example, Richardson-Self, 2018) also carries the risk of trivializing misogynist hate speech in the understanding of the mainstream (Sponholz, 2021: 22). Thus, it makes sense to compare online bystander intervention in misogynist online hate speech with that in racist and homophobic online hate speech.

Research based on the BIM suggests that, in order for them to intervene, bystanders must initially perceive hate speech online as threatening or harmful to the targeted group (Leonhard et al., 2018). In this study, we go a step further. We assume that it is already sufficient to perceive online hate speech as norm-transgressing, and thus as uncivil communication, to boost the bystander intervention process (Naab et al., 2018). Thus, it is first necessary to investigate the extent to which bystanders perceive hate speech against LGBTQIA+, individuals with a migration background, or women as having different degrees of incivility. Therefore, we pose the following research question:

> **RQ1:** To what extent does online hate speech against LGBTQIA+, against people with a migration background, and against women differ in terms of perceived incivility?

According to the BIM (Latané and Darley, 1970), bystanders to hate speech are more willing to intervene by uttering counterspeech the more threatening or norm-transgressing they perceive the incident to be and the more they feel personally responsible to intervene (Leonhard et al., 2018; Naab et al., 2018; Ziegele et al., 2020). Therefore, we expect the intentions of bystanders to intervene by using counterspeech in reaction to hate speech against different social groups to be mediated accordingly. Hence, we assume:

> **H1:** The willingness to utter counterspeech after reading online hate speech against LGBTQIA+, individuals with a migration background, or women is mediated by the perceived incivility of the online hate speech and, in turn, the perceived personal responsibility to intervene.

Figure 1 illustrates the hypothesized model.

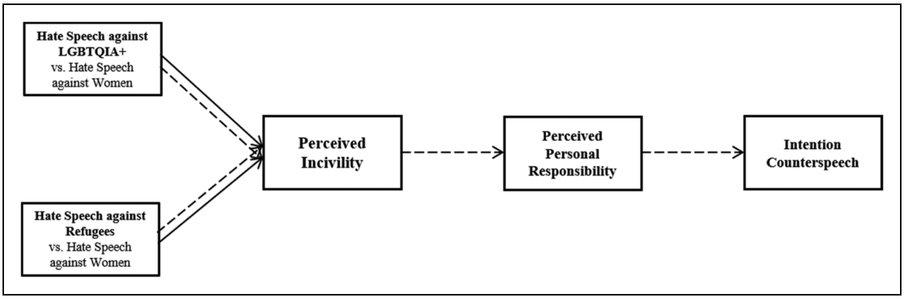

Hate Speech against LGBTQIA+ vs. Hate Speech against Women
Hate Speech against Refugees vs. Hate Speech against Women
Perceived Incivility
Perceived Personal Responsibility
Intention Counterspeech

Figure 1. Hypothesized model.

Materials and methods

Design and participants

To test our hypothesis and research question, we conducted an online experiment with three experimental conditions involving hate speech against different social groups in a 3 × 1 between-subjects factorial design. Thus, we varied whether participants were confronted with (1) hate speech against LGBTQIA+ (homophobia), (2) hate speech against individuals with a migration background (racism), or (3) hate speech against women (misogyny). Because the study investigates hate speech on social media platforms, participants were recruited via the most heterogeneous selection of social networking sites (Facebook, Instagram) and messenger apps (WhatsApp). On these platforms, we focused on various university study programs and sites whose aim is to advertise and invite participation in scientific studies. For the present study, participation was voluntary and unpaid. In terms of ethical considerations, for example, respondents were assured of the anonymity of their data, given a detailed debriefing about the study's research interest and the fictional stimulus material, and provided with contact information in case of any questions or remarks. A total of 140 online users took part and were randomly assigned to one of the experimental groups. From the total sample, we excluded participants who completed the questionnaire in less than half the median time for all participants ($n = 10$). After that, 130 participants remained in the final sample (60% female, mean age = 23 years, $SD = 5.48$, higher education: 95%). Therefore, the sample has slightly more women, is younger, and is highly educated compared to the German population, which must be considered when interpreting the absolute findings. However, experimental groups did not differ significantly with regard to gender, $\chi^2(2) = 1.05$, $p = .59$, age, $F(2, 123) = 0.12$, $p = .88$, and education, $\chi^2(2) = 4.45$, $p = .11$. Thus, these possibly confounding variables are similarly distributed across the experimental groups, and systematic biases are less likely.

Stimulus materials

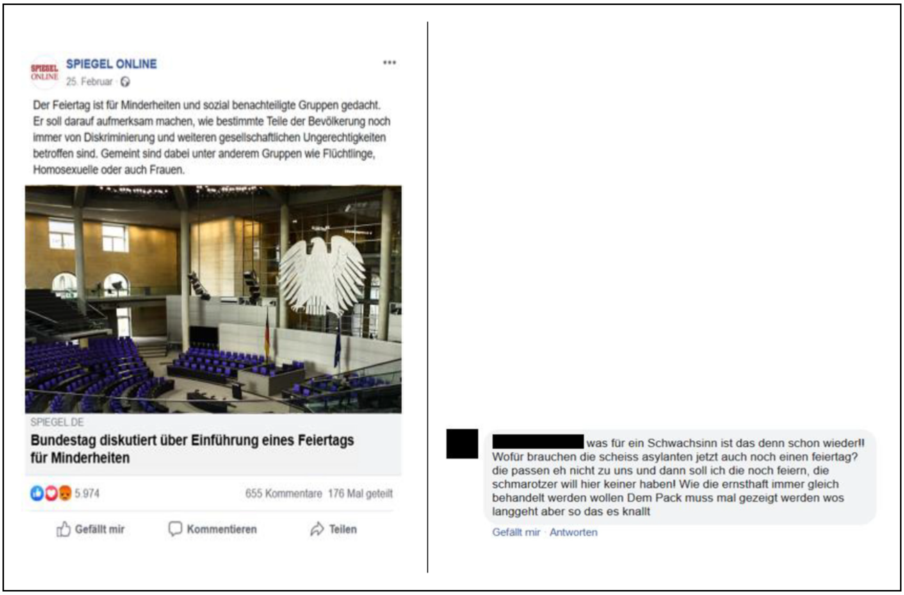
Participants were asked to read a (fictitious) post on a screenshot of a Facebook thread from the German news website Spiegel Online, which is considered a high quality and comparatively trustworthy media brand (Newman et al., 2021) and is highly frequented.

The post included a teaser to an article entitled ‘Parliament Discusses Introduction of a Holiday for Minorities’ and referred to a potential new holiday dedicated to discriminated minorities. In order to distinguish the three target groups of online hate speech, the holiday should be dedicated to either refugees; to a specific group of people with a migration background, which we chose because online hate speech against them is particularly common in Germany (Geschke et al., 2019; Nennstiel and Isenberg, 2022); to people of the LGBTQIA+ community; or to women. The preview of the linked article was identical for all experimental groups. Below the posting, a user comment containing hate speech was displayed. The stimulus appeared in the Facebook design and had the same number of reactions, comments, and shares as a comparable real post.

The accompanied user comment served as the independent variable, which included hateful expressions against LGBTQIA+, refugees, or women (see Figure 2). The comment had characteristics found to be typical of hate speech, namely derogations and discriminations of the specific social groups (‘how they seriously always want to be treated equally’), name-calling and vulgarities, and an explicit incitement to violence (‘you have to show the *** what’s what, but in a way that makes it pop’) (Hawdon et al., 2017; Kenski et al., 2020; Stryker et al., 2016). Table 1 shows the exact wording for the three experimental groups. The questionnaire concluded with a detailed debriefing.

Measures

As a short treatment check, participants were asked to recall the *target group* at which the hate speech was directed. Afterward, we measured the *perceived incivility* of the user



SPIEGEL ONLINE

25. Februar

Der Feiertag ist für Minderheiten und sozial benachteiligte Gruppen gedacht. Er soll darauf aufmerksam machen, wie bestimmte Teile der Bevölkerung noch immer von Diskriminierung und weiteren gesellschaftlichen Ungerechtigkeiten betroffen sind. Gemeint sind dabei unter anderem Gruppen wie Flüchtlinge, Homosexuelle oder auch Frauen.

SPIEGEL.DE

Bundestag diskutiert über Einführung eines Feiertags für Minderheiten

[redacted]

[redacted] was für ein Schwachsinn ist das denn schon wieder!! Wofür brauchen die scheiss asylanten jetzt auch noch einen feiertag? die passen eh nicht zu uns und dann soll ich die noch feiern, die schmarotzer will hier keiner haben! Wie die ernsthaft immer gleich behandelt werden wollen Dem Pack muss mal gezeigt werden wos langgeht aber so das es knallt

Figure 2. Example of the German stimulus material (experimental group of hate speech against refugees).

Table 1. Stimulus materials.

	Hate speech against …	
	English translation	German original
Refugees	What kind of bullshit is this again!! For what do the damn asylum seekers also need a holiday? they don't fit to us anyway and then I should celebrate them, nobody wants to have the parasites here! How they seriously always want to be treated equally. You have to show the pack what's what, but in a way that makes it pop	was für ein Schwachsinn ist das denn schon wieder!! Wofür brauchen die scheiss asylanten jetzt auch noch einen feiertag? die passen eh nicht zu uns und dann soll ich die noch feiern, die schmarotzer will hier keiner haben! Wie die ernsthaft immer gleich behandelt werden wollen. Dem Pack muss mal gezeigt werden wos langgeht aber so das es knallt
LGBTQIA+	What kind of bullshit is this again!! For what do the damn homos also need a holiday? now the pansies here are already allowed to marry and then I am forced to celebrate that. What they do should be prohibited! How they seriously always want to be treated equally. You have to show the pack what's what, but in a way that makes it pop	was für ein Schwachsinn ist das denn schon wieder!! Wofür brauchen die scheiss homos jetzt auch noch einen feiertag? jetzt dürfen die schwuchtel hier schon heiraten und dann soll ich das noch feiern, was die da machen gehört doch verboten! Wie die ernsthaft immer gleich behandelt werden wollen. Dem Pack muss mal gezeigt werden wos langgeht aber so das es knallt
Women	What kind of bullshit is this again!! For what do the damn bitches also need a holiday? also stop me with this same payment garbage, what men and women do is never comparable! How they seriously always want to be treated equally. You have to show the pack what's what, but in a way that makes it pop	was für ein Schwachsinn ist das denn schon wieder!! Wofür brauchen die scheiss weiber jetzt auch noch einen feiertag? hört mir auch auf mit diesem gleiche bezahlungs müll, was männer und frauen leisten ist niemals vergleichbar! Wie die ernsthaft immer gleich behandelt werden wollen. Dem Pack muss mal gezeigt werden wos langgeht aber so das es knallt

Notes. All stimulus materials were in German in the original; for a better understanding, we also provide the English translation.

comment on a seven-point semantic differential (1 = 'civil' and 7 = 'uncivil,' $M = 6.80$, $SD = 0.56$) (Kenski et al., 2020).

Perceived *personal responsibility* was assessed by inquiring about participants' agreement with the following items (five-point scale; 1 = 'does not apply at all' and 5 = 'fully applies'): 'I feel personally responsible for supporting the affected group'; 'I consider it my duty to help the affected group'; and 'It is my obligation to do something about this comment' ($\alpha = .88$, $M = 3.19$, $SD = 1.08$) (Leonhard et al., 2018).

The *intention of bystander intervention* was evaluated using two items based on a list of strategies used to counter online hate speech (Bartlett and Krasodomski-Jones, 2015;

Leonhard et al., 2018). Participants rated their behavioral intention with the following items (five-point scale; 1 = ‘does not apply at all’ and 5 = ‘fully applies’): ‘I write a comment myself that contradicts the other comment’ and ‘I call upon my friends and acquaintances to comment negatively on the statement’ ($r = .55$, $p < .01$, $M = 1.75$, $SD = 0.84$).

Results

Treatment check

To determine whether our treatment worked adequately, we asked participants if they could recall the target group of the hate comment they read. The treatment check revealed that the vast majority of participants were able to name the target group correctly (hate speech against LGBTQIA+: 98%; hate speech against refugees: 98%; hate speech against women: 95%).

Direct effects of online hate speech against different social groups on perceived incivility

To answer our research question (*RQ1*), we calculated a univariate analysis of variance to test how uncivil the forms of hate speech against different social groups were perceived to be and to what extent they differed in the degree of incivility as perceived by the participants. The analysis revealed that, although all three forms of hate speech were perceived to be somewhat uncivil as they were rated above the scale’s midpoint, hate speech against women was estimated to be the most uncivil ($M = 6.93$, $SD = 0.35$), followed by hate speech against refugees ($M = 6.82$, $SD = 0.45$), and hate speech against members of the LGBTQIA+ community ($M = 6.66$, $SD = 0.78$), $F(2, 126) = 2.49$, $p = .09$, $\eta^2_{part} = .04$.

Thus, the three forms of hate speech are perceived as having only marginally different levels of incivility, with misogyny and homophobia differing the most. This may be because the sample comprises participants who self-identify as heterosexual and cisgender, which could promote their perception of incivility with regard to sexist or, in the case of women, misogynistic online hate speech. This should be considered when interpreting the findings. According to this result, we created two dummy-coded variables of the experimental groups for the follow-up analyses to test our hypothesis that misogyny was the reference category for both homophobia and racism. By taking misogyny as a reference category, we are also taking into consideration that hates speech against women is treated differently from a legal perspective, resulting in the risk of misogyny being interpreted as less inappropriate. This allows us to determine whether it is also reflected in bystander intervention.

Indirect effects of online hate speech against different social groups

To test our mediation hypothesis (*H1*), we used structural equation modeling (SEM) with maximum likelihood estimation (ML) utilizing Mplus (Muthén and Muthén, 2010). Bootstrap standard errors and bias-corrected 95% confidence intervals were generated based on 10,000 bootstrap samples. The treatment was entered as dummy variables in the model using hate speech against women as a reference category, as reported above. Due to their different measuring scales, all dependent variables were z-transformed before including them in the SEM. Zero-order correlations of all constructs represented in the model are presented in Table 2. The model offered a good fit for the data, $\chi^2(16) = 22.09$, $p = .14$, $CFI = .98$, $RMSEA = .05$, $p = .41$, $SRMR = .03$ (Hu and Bentler, 1999). Figure 3 illustrates the results of the SEM. Beyond that, the direct associations represented in the SEM are demonstrated in Table 3.

Using this model, we tested our assumption that hate speech against both LGBTQIA+ and individuals with a migration background (compared to hate speech against women) increases the intention to provide counterspeech, mediated by the perception of incivility and, in turn, a stronger feeling of personal responsibility. First, we consider the direct associations in the SEM between the variables that are relevant to our presumed mediation. Compared to hate speech containing misogynist devaluations, hate speech

Table 2. Zero-order correlations.

	1	2	3	4	5
1. Hate speech against LGBTQIA+[a]	1				
2. Hate speech against refugees[a]	−.52***	1			
3. Perceived incivility	−.18*	.03	1		
4. Perceived personal responsibility	−.08	.02	.21*	1	
5. Intention counterspeech	.01	−.001	.03	.41***	1

Notes. $N = 128$, [a]Hate speech against women is reference group, $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

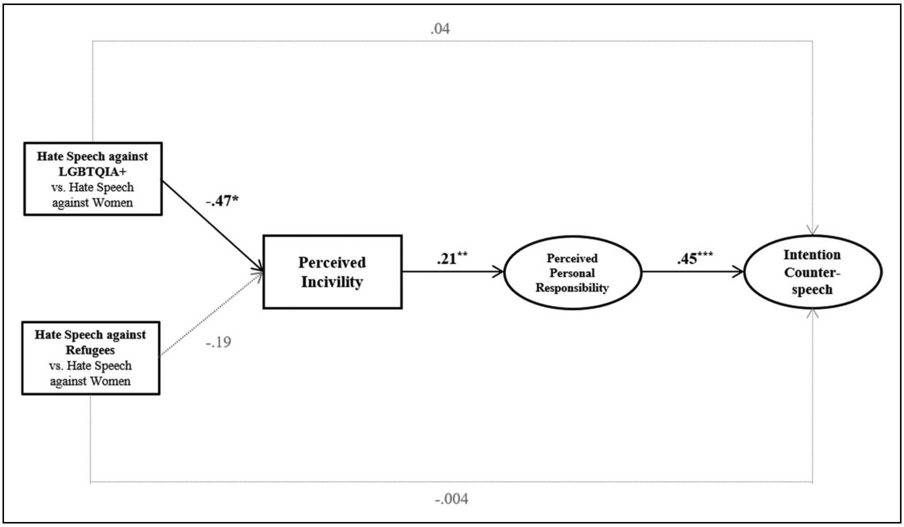

.04
Hate Speech against LGBTQIA+ vs. Hate Speech against Women
-.47*
Perceived Incivility
.21**
Perceived Personal Responsibility
.45***
Intention Counter-speech
Hate Speech against Refugees vs. Hate Speech against Women
-.19
-.004

Figure 3. Effects of online hate speech against different social groups on bystander intervention (Misogyny as reference category).

Notes. $N = 129$, $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$, unstandardized coefficients, significance-testing via bootstrap method (10,000 samples), 95% bias-corrected boot-strap confidence intervals (*CI*), $\chi^2(16) = 22.09$, $p = .14$, $CFI = .98$, $TLI = .97$, $RMSEA = .05$, $p = .41$, $SRMR = .03$.

Table 3. Direct effects of online hate speech against different social groups on bystander intervention (misogyny as reference category).

	Perceived incivility			Perceived personal responsibility			Intention counterspeech		
Predictor	β	B	SE	β	B	SE	β	B	SE
Hate speech against LGBTQIA+[a]	−.23*	−0.47*	0.23				.02	0.04	0.24
Hate speech against refugees[a]	−.09	−0.19	0.15				−.002	−0.004	0.23
Perceived incivility				.23**	0.21**	0.07			
Perceived personal responsibility							.49***	0.45***	0.10

Notes. $N = 129$, [a]Hate speech against women is reference group, STDY standardization (Muthén and Muthén, 2010), ß = standardized coefficients, B = unstandardized coefficients, SE = standard errors for B, *$p < .05$, **$p < .01$, ***$p < .001$, $\chi^2(16) = 22.09$, $p = .14$, $CFI = .98$, $TLI = .97$, $RMSEA = .05$, $p = .41$, $SRMR = .03$.

against LGBTQIA+ led to significantly less perceived incivility, $B = -0.47$, $SE = 0.23$, $p = .04$. However, racist hate comments were not perceived to be less uncivil compared to misogynistic hate comments, $B = -0.19$, $SE = 0.15$, $p = .21$. In addition, neither reading homophobic hate speech compared to misogynistic hate speech, $B = 0.04$, $SE = 0.24$, $p = .86$, nor reading racist hate speech instead of misogynistic hate speech, $B = -0.004$, $SE = 0.23$, $p = .99$, affected the intention to intervene with counterspeech directly. In line with the BIM, our SEM showed that the degree of perceived incivility positively affected participants' feelings of personal responsibility, $B = 0.21$, $SE = 0.07$, $p = .002$. In turn, feeling personally responsible to intervene boosted participants' intention to provide counterspeech in reaction to hate speech, $B = 0.45$, $SE = 0.10$, $p < .001$.

Therefore, there was, indeed, a significant negative indirect effect of homophobic hate speech on participants' intentions to intervene by counterarguing mediated through perceived incivility and, in turn, the feeling of personal responsibility, $B_{\text{ind\_h}} = -0.04$, $SE = 0.02$, $CI = [-0.11, -0.01]$. Yet there was no indirect effect of racist hate speech on the intention to utter counterspeech mediated through the perceived incivility and the feeling of personal responsibility in turn, $B_{\text{ind\_r}} = -0.02$, $SE = 0.02$, $CI = [-0.07, 0.01]$. In summary, the hypothesized mediation in *H1* was confirmed although only for misogynistic online hate speech compared to homophobic online hate speech.

Discussion

Relatively few users actively post uncivil comments on digital media, while a large majority are uninvolved readers or, in the case of online hate speech, bystanders (Nennstiel and Isenberg, 2022). Therefore, this group of bystanders is central to combatting online hate speech, for example, through counterspeech. This justifies the study of mechanisms of bystander intervention in online hate speech. Using the bystander intervention model (Latané and Darley, 1970), we examined the sequential mechanisms by which bystanders respond to hate speech against various social groups, namely (1) by perceiving the hate speech to be uncivil and (2) by feeling personally responsible to intervene. In doing so, we complement the existing research on bystander intervention (based on the BIM) in the context of online hate speech in two aspects. First, we demonstrate that perceiving an utterance of hate speech as uncivil may trigger the bystander intervention process. In line with previous research (Leonhard et al., 2018), bystanders of hate speech did not automatically intervene after reading hate speech. However, the greater the degree to which they perceived the hate speech as uncivil, the more they felt personally responsible to intervene. This feeling of personal responsibility, in turn, led to bystander intervention by counterarguing and/or encouraging others to do so. Thus, even hate speech that at present is not prosecuted by law (for example, hate speech against women) is perceived as uncivil, and consequently, the perception of responsibility and the intention to intervene arise. This is interesting in light of the fact that studies of bystander intervention in uncivil behavior on digital media have, thus far, often focused on the perceived threat of the incidents. Thus, the perception that norms of discourse in a democratic society are being transgressed may be sufficient to trigger the process of counterspeech. This is important because stand-alone online hate speech can create the impression among bystanders that these verbal disparagements are appropriate and tolerated in a democratic society (Kümpel and Rieger, 2019; see also Iganski, 2020). Moreover, it is conceivable that such an impression may reinforce potential negative social and psychological consequences for affected communities (Keel et al., 2022; Leets, 2002). In addition, it is certainly necessary to examine the form intervening users may employ to express counterspeech, which should be investigated in follow-up studies because hateful counterspeech, in particular, may lead to even more hostile responses, further damaging civil discourse (Chen and Lu, 2017).

In addition, while personal forms of uncivil online communication (such as insults or name-calling) violate norms of polite discourse, public forms transgress norms of deliberative and democratic discourse (Papacharissi, 2004). Online hate speech, which is always directed against social-group identities, can in principle contain both (Schwertberger and Rieger, 2021). Thus, it would be interesting for subsequent studies to determine which content-related characteristics make it more likely that an incident of hate speech is perceived as uncivil, that is, as transgressing norms of politeness and/or discursive norms, or even threatening to bystanders' own social identity. It would then be necessary to examine the extent to which these assessments have different effects on the willingness to intervene (for example, by engaging in counterspeech or reporting such posts to service providers).

Second, to the best of our knowledge, the present study is one of the first to investigate not only the mechanisms of bystander intervention in online hate speech but also to systematically compare bystander intervention in hate speech against different social groups. By focusing on three frequently targeted social groups – namely members of the LGBTQIA+ community; individuals with a migration background, in general, and refugees specifically; and women – we found that all three forms of hate speech were perceived to be uncivil utterances, as indicated by bystanders' agreement in each case above the scale's midpoint. Yet an analysis of variance comparing homophobic, racist, and misogynistic hate speech indicated only marginal differences with regard to perceived incivility. This means that similarly serious forms of online hate speech are not perceived as essentially different in regard to lack of civility and, thus, normatively inappropriate in a democratic society. From a democratic perspective, glaring differences here would also be worrisome because they would suggest that hate speech against some social groups is seen as more acceptable than hate speech against others. However, while racist and misogynistic hate speech did not differ regarding perceived incivility, structural equation modeling suggested that hate speech against LGBTQIA+ was perceived to be slightly less uncivil compared to hate speech against women. A reason for this could be that our sample comprised mostly people who identify as heterosexual and, thereby, are more likely to be affected by sexist (and, more specifically, misogynistic) hate speech themselves and perceive it as more inappropriate, which we discuss in greater detail below.

Thus, in line with existing research on bystander intervention using the BIM (Leonhard et al., 2018), we identified a weak negative indirect effect of homophobic hate speech as opposed to misogynistic hate speech on the intention to counterargue. However, there was no indirect effect comparing racist hate speech to misogynistic hate speech. Hence, compared to homophobic hate speech, participants perceived misogynistic hate speech as more uncivil, which increased their feelings of personal responsibility to intervene and, in turn, boosted their willingness to counterargue. Accordingly, we were able to successfully demonstrate the steps underlying the BIM for two kinds of hate speech: homophobia and misogyny (but not for racism compared to misogyny). This indicates that bystanders are more likely to help the social group that they believe is subjected to more uncivil verbal abuse. This can certainly be desirable for alleviating negative consequences for those affected; however, it is important not to lose sight of the fact that others may also be affected by online hate speech that is perceived as less serious but can be just as damaging to those affected and where solidarity-based intervention is necessary. This should be considered and communicated in appropriate initiatives to promote online counterspeech.

It is important to keep in mind that our sample had a majority of women, that is, 60%, directly affected by misogynistic hate speech. Especially if the severity of the hate speech is moderate, there is a difference in its perception based on social identification with the targeted group, such as identifying as a woman (Wojatzki et al., 2018). In contrast, in our sample, none of the participants identified as gay (bi- or pansexual: 7%), and only 16% indicated having at least one parent with a migration background. However, in general, people are believed to react more sensitively to hate speech that targets their own relevant social group. This could be addressed in future studies by further differentiating the extent to which identification with the affected group condition perceived incivility and, ultimately, bystander intervention.

Our study has several limitations. First, we used a highly educated convenience sample, which could have led to an overestimation of the effects we found. Therefore, follow-up studies should investigate the perception of hate speech against different social groups using more heterogeneous or even representative samples in order to be able to provide statements about the absolute level of perceived incivility.

Second, we decided to compare hate speech against two minority groups (LGBTQIA+ and individuals with a migration background) and a social group (women) that are relatively frequent targets of such utterances in digital media. In doing so, we have inevitably neglected hate speech against other groups, such as social groups characterized by certain attitudes and beliefs (for example, related to politics, religion, sustainable lifestyles), disabilities and chronic illnesses, or professional characteristics (for example, politicians, journalists) or those affected by such utterances with varying frequency (Geschke et al., 2019). Additionally, we cannot make a statement about the extent to which all three forms differ from a neutral comment in terms of perceived incivility. Follow-up studies should, therefore, extend the perceived differences regarding the incivility of hate speech against different groups and compare them with a neutral control comment in order to make predictions about bystander intervention here.

Third, we measured bystanders' intentions to intervene by asking them about the extent to which they were willing to counterargue or motivate others instead of actually enabling them to do so. Therefore, future studies could investigate bystander intervention in different forms of online hate speech with measurements of higher external validity, for example, by allowing participants to write comments or to react with emojis or likes, or by enabling them to flag hateful user comments. Fourth, our results regarding the indirect effects of hate speech against different social groups were shaped using misogynistic hate speech as the reference category for both homophobic and racist hate speech. Among other reasons, because women are not considered a minority by German law and these groups differentiated most clearly in terms of perceived incivility, this approach seemed plausible to us, but different coding might have led to different interpretations.

To summarize, we extend the existing literature on BIM in the case of hate speech by showing that hate speech is perceived as uncivil in different ways depending on the targeted social group and may be perceived as having varying degrees of urgency to intervene from the perspective of bystanders. Hence, these findings stress that one way to promote the perceived responsibility of bystanders and, ultimately, their courageous intervention in hate speech is to make it clear to bystanders that even implicit, seemingly harmless incidents of hate speech can have serious consequences for those affected and that—regardless of the social group attacked—intervention is essential.

Acknowledgement

The authors would like to thank Christian Balser for his valuable input and his support in implementing the study.

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
Ursula Kristin Schmid https://orcid.org/0000-0002-1892-002X
Diana Rieger https://orcid.org/0000-0002-2417-0480

References

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

Bastiaensens S, Vandebosch H, Poels K, et al. (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.

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.

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

Döring N and Mohseni MR (2019) Male dominance and sexism on YouTube: results of three content analyses. *Feminist Media Studies* 19(4): 512–524.

Döring N and Mohseni MR (2020) Gendered hate speech in YouTube and YouNow comments: RESULTS of two content analyses. *Studies in Communication and Media* 9(1): 62–88.

Fasoli F, Paladino MP, Carnaghi A, et al. (2016) Not 'just words': exposure to homophobic epithets leads to dehumanizing and physical distancing from gay men. *European Journal of Social Psychology* 46(2): 237–248.

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

Hawdon J, Oksanen A and Räsänen P (2017) Exposure to online hate in four nations. A cross-national consideration. *Deviant Behavior* 38(3): 254–266.

Henry N and Powell A (2018) Technology-facilitated sexual violence: a literature review of empirical research. *Trauma, Violence, & Abuse* 19(2): 195–208.

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): 1–55.

Iganski P (2020) Civil courage as a communicative act: countering the harms of hate violence. *Pragmatics and Society* 11(2): 316–335.

Isenberg M (2019) *Hate Speech. Zentrale Untersuchungsergebnisse der aktuellen forsa-Studie* 2019. Landesanstalt für Medien NRW. URL (accessed 21 July 2021): https://www.medienanstalt-nrw.de/fileadmin/user_upload/lfm-nrw/Service/Pressemitteilungen/Dokumente/2019/forsa_LFMNRW_Hassrede2019_Ergebnispraesentation.pdf.

Keel C, Wickes R and Benier K (2022) The vicarious effects of hate: inter-ethnic hate crime in the neighborhood and its consequences for exclusion and anticipated rejection. *Ethnic and Racial Studies* 45(7): 1283–1303.

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

Kenski K, Coe K and Rains SA (2020) Perceptions of uncivil discourse online: an examination of types and predictors. *Communication Research* 47(6): 795–814.

Koehler C and Weber M (2018) 'Do I really need to help?!' Perceived severity of cyberbullying, victim blaming, and bystanders' willingness to help the victim. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* 12(4): article 4.

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

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

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

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.

LfM (2022) *Hate Speech Forsa-Studie 2022*. URL (accessed 2 February 2023): https://www.medienanstaltnrw.de/fileadmin/user_upload/NeueWebsite_0120/Themen/Hass/LFM_Hatespeech_forsa_2022_01.pdf.

Lück J and Nardi C (2019) Incivility in user comments on online news articles: investigating the role of opinion dissonance for the effects of incivility on attitudes, emotions and the willingness to participate. *Studies in Communication and Media* 8(3): 311–337.

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

McCoy SK and Major B (2003) Group identification moderates emotional responses to perceived prejudice. *Personality and Social Psychology Bulletin* 29(8): 1005–1017.

Meleagrou-Hitchens A and Kaderbhai N (2017) *Perspectives on online radicalization. Literature review 2006*-2016. Vox Pol. URL (accessed 21 April 2021): https://icsr.info/wp-content/uploads/2017/05/ICSR-Paper_Research-Perspectives-on-Online-Radicalisation-A-Literature-Review-2006-2016.pdf.

Mohseni MR (2021) Sexistische online-Hassrede auf Videoplattformen. In: Wachs S, Koch-Priewe B and Zick A (eds) *Hate Speech – Multidisziplinäre Analysen und Handlungsoptionen*. Wiesbaden: Springer VS, pp.39–51.

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

Naab TK, Kalch A and 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.

Nennstiel S and Isenberg M (2022) Hate speech forsa-Studie 2022. Zentrale Untersuchungsergebnisse. URL (accessed 5 May 2022): https://www.medienanstalt-nrw.de/themen/hass/forsa-befragung-zur-wahrnehmung-von-hassrede.html.

Newman N, Fletcher R, Schulz A, et al. (2021) Reuters institute digital news report 2021. URL (accessed 9 May 2022): https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2021-06/Digital_News_Report_2021_FINAL.pdf.

Nielsen LB (2002) Subtle, pervasive, harmful: racist and sexist remarks in public as hate speech. *Journal of Social Issues* 58(2): 265–280.

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.

Papacharissi Z (2004) Democracy online: civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society* 6(2): 259–283.

Preuß M, Tetzlaff F and Zick A (2017) *'Publizieren wird zur Mutprobe.' Studie zur Wahrnehmung von und Erfahrung mit Angriffen unter JournalistInnen*. URL (accessed 21 April 2021): https://mediendienst-integration.de/fileadmin/Dateien/Studie-hatespeech.pdf.

Richardson-Self L (2018) Woman-hating: on misogyny, sexism, and hate speech. *Hypatia* 33(2): 256–272.

Schieb C and Preuss M (2016) Governing hate speech by means of counterspeech on Facebook. In: 66th Annual Conference of the International Communication Association, Fukuoka, Japan, 9–13 June 2016.

Schwertberger U and Rieger D (2021) Hass und seine vielen Gesichter: Eine sozial- und kommunikationswissenschaftliche Einordnung von hate speech. In: Wachs S, Koch-Priewe B and Zick A (eds) *Hate Speech – Multidisziplinäre Analysen und Handlungsoptionen*. Wiesbaden: Springer VS, pp.52–77.

Sękowska-Kozłowska K, Baranowska G and Gliszczyńska-Grabias A (2022) Sexist hate speech and the International Human Rights Law: towards legal recognition of the phenomenon by the United Nations and the council of Europe. *International Journal for the Semiotics of Law - Revue Internationale de Sémiotique Juridique* 35: 2323–2345.

Sellers RM, Caldwell CH, Schmeelk-Cone KH, et al. (2003) Racial identity, racial discrimination, perceived stress, and psychological distress among African American young adults. *Journal of Health and Social Behavior* 44(3): 302–317.

Sponholz L (2021) Hass mit likes: hate speech als Kommunikationsform in den social media. In: Wachs S, Koch-Priewe B and Zick A (eds) *Hate Speech – Multidisziplinäre Analysen und Handlungsoptionen*. Wiesbaden: Springer VS, pp.15–37.

Steinl L (2019) Hasskriminalität und geschlechtsbezogene Gewalt gegen Frauen: Eine Einführung aus strafrechtlicher Perspektive. *Zeitschrift für Rechtssoziologie* 38(2): 179–207.

Stryker R, Conway BA and Danielson JT (2016) What is political incivility? *Communication Monographs* 83(4): 535–556.

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: 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.

Wachs S, Schubarth W, Krause N, et al. (2021) Hate speech als Herausforderung für Schule und Lehrkräftebildung. In: Wachs S, Koch-Priewe B and Zick A (eds) *Hate Speech – Multidisziplinäre Analysen und Handlungsoptionen*. Wiesbaden: Springer VS, pp.279–297.

Wilhelm C and Joeckel S (2019) Gendered morality and backlash effects in online discussions: an experimental study on how users respond to hate speech comments against women and sexual minorities. *Sex Roles* 80(7): 381–392.

Wilhelm C, Joeckel S and Ziegler I (2020) Reporting hate comments: investigating the effects of deviance characteristics, neutralization strategies, and users' moral orientation. *Communication Research* 47(6): 921–944.

Wojatzki M, Horsmann T, Gold D, et al. (2018) Do women perceive hate differently: examining the relationship between hate speech, gender, and agreement judgments. In: Paper Presented at the 14th Conference on Natural Language Processing (KONVENS 2018), Vienna, Austria.

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, Naab TK and Jost P (2020) Lonely together? Identifying the determinants of collective corrective action against uncivil comments. *New Media & Society* 22: 731–751.

Ziegele M, Viehmann C and Weber M (2018) Socially destructive? Effects of negative and hateful user comments on readers' donation behavior toward refugees and homeless persons. *Journal of Broadcasting & Electronic Media* 62(4): 636–653.