Computers in Human Behavior 104 (2020) 106192

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: http://www.elsevier.com/locate/comphumbeh

Online Hate Does Not Stay Online – How Implicit and Explicit Attitudes Mediate the Effect of Civil Negativity and Hate in User Comments on Prosocial Behavior

Mathias Weber [b,*], Christina Viehmann [b], Marc Ziegele [c], Christian Schemer [b]

[b] Department of Communication, University of Mainz, Germany
[c] Department of Social Sciences, University of Duesseldorf, Germany

ARTICLE INFO

Keywords:
Incivility
Hate speech
User comments
Implicit attitudes
Behavioral effects

ABSTRACT

Incivility and hateful language in user comments are met with growing concern among politicians, the general public, and scholars. There are fears that such comments may decrease social cohesion and ultimately result in less prosocial behavior among citizens. We investigate whether hate, or even civil negativity in user comments alone, inhibit actual prosocial behavior through recipients' explicit and implicit attitudes. In an online experiment, 253 participants read user comments (neutral, civil-negative, hateful) about refugees and received five Euros which they could donate for a refugee aid organization or keep for themselves. The results show that participants confronted with hateful or negative user comments donated less money. The effect was mediated by both explicit and implicit attitudes toward refugees with hate having a stronger influence via implicit attitudes. The results are discussed in light of possible measures for reducing the behavioral impact of negative and hateful user comments.

Online Hate Does Not Stay Online – How implicit and Explicit Attitudes Mediate the Effect of Negativity and Hate in User Comments on Prosocial Behavior.

User commenting has become a prominent form of lay engagement with journalistic content. While only a minority of internet users posts comments to online news articles regularly (e.g., Japan: 6%, Germany: 8% US: 24%, Turkey: 31%; Newman, Fletcher, Levy, & Nielsen, 2016), a considerable share reads these comments at least occasionally (US: 49%, Germany: 63%; Köcher, 2016; Stroud, Van Duyn, & Peacock, 2016). Moreover, user comments have become a topic of public concern. Although most users express their thoughts in a civil manner, a significant number of comments contains aggressive, derogatory, and disrespectful statements. In fact, recent research has shown that between 22 and 33 percent of the comments posted on the websites and Facebook sites of various local and national newspapers contained disparaging language, profanity, or unsubstantiated accusation of disingenuousness (Coe, Kenski, & Rains, 2014; Su et al., 2018). In extreme cases, such incivility in comment sections culminates in hate speech, encompassing verbal aggression against an individual or a group solely based on gender, race, religion, sexual orientation, or other social category affiliations (Gagliardone, Gal, Alves, & Martinez, 2015). On the Facebook pages of national news outlets, such extreme forms of incivility occur far less frequently than other forms of incivility (seven percent, Su et al., 2018), but on the Facebook pages of local news outlets, extreme incivility seems even more widespread than less severe forms (Su et al., 2018). Additionally, more than two thirds of German online users have encountered hate speech in comment sections at least once (LfM, 2017). Some news outlets have therefore closed their comment sections or restricted the possibility to post comments to soft news topics only (e.g., Stroud, Scacco, Muddiman, & Curry, 2015). In Germany, the government additionally passed a law to tame the apparent surge in hateful user comments; this "Network Enforcement Act" defines strict rules for handling hate speech in postings that social media platforms now need to enforce (BMJV, 2017).

The rationale behind such remedies is the notion that uncivil or hateful user comments might have undesirable effects on readers; that they could provoke anti-social attitudes and ultimately lead to less prosocial or more aggressive behavior (e.g., Greenblatt, 2015; Lobo, 2015). This has especially been argued with respect to hateful comments directed at ethnic minorities, migrants, and refugees (e.g., Gagliardone

[*] Corresponding author.
E-mail addresses: mathias.weber@uni-mainz.de (M. Weber), christina.viehmann@uni-mainz.de (C. Viehmann), ziegele@hhu.de (M. Ziegele), schemer@uni-mainz.de (C. Schemer).

https://doi.org/10.1016/j.chb.2019.106192
Received 6 June 2019; Received in revised form 5 November 2019; Accepted 7 November 2019
Available online 8 November 2019

et al., 2015). The current research aims at empirically substantiating such claims by investigating the attitudinal and behavioral effects of uncivil or hateful user comments toward migrants and refugees.

Various studies have suggested that reading user comments can result in altered perceptions of issues, negative emotions, attitude change, and even in changes in readers' commenting behavior (e.g., Flemming, Cress, and Kimmerle, 2017; Gervais, 2014; Rösner, Winter, and Krämer, 2016). For example, Hsueh, Yogeeswaran, and Malinen (2015) found that user comments containing strong prejudice against Asian-Americans led to more prejudice amongst subjects exposed to these comments. However, extant studies investigating attitudinal or behavioral effects of user comments come with several limitations, which the current study will address:

(1) As yet, the measurement of behavioral effects is limited to writing comments in mock-up online environments (for an exception, see Ziegele, Koehler, & Weber (2018)). The current study aims at examining whether exposure to hate speech can inhibit real and more consequential pro-social acts, such as donating money for social grous that are stigmatized in hate speech comments.

(2) Most studies assess attitudes and behavioral intentions relying on explicit measures (for an exception, see Hsueh et al., 2015). However, individuals often cannot or do not want to verbalize their attitudes and intentions, because they are not sufficiently aware of them or because they perceive them to be socially undesirable (Greenwald, Poehlman, Uhlmann, & Banaji, 2009). Even more so: Although an attitude might be activated and salient in a person's mind, she or he might not explicitly express it because the person consciously refuses to applicate it (Gawronski & Bodenhausen, 2006). Therefore, the present study does not only rely on explicit measures of attitudes, but extends the perspective to implicit measures as well. These explicit and implicit attitudes are then modelled as potential mediators of behavioral effects of user comments.

(3) Only few studies have distinguished effects of negative, yet civil comments (i.e., comments expressing opposition or disagreement in a civil manner) from those of uncivil and hateful comments (Chen & Lu, 2017). Such a differentiation is crucial to assess whether (civil) disagreement alone causes detrimental effects of comments or whether the disagreement needs to be expressed in an uncivil manner. This study will, hence, compare the cognitive and behavioral effects of hateful and negative, yet civil user comments with those of neutral comments.

(4) Finally, research focusing on user comments is mostly conducted with convenience and student samples (e.g., Flemming et al., 2017; Hsueh et al., 2015; Lee & Jang, 2010). Given that attitudes and behavioral tendencies vary with social background, the current research will draw on a more diverse sample, that is, a sample that matches the population distribution regarding gender, age, education, and home region.

In sum, this study draws on a diverse sample to investigate the differential effects of negative-civil and hateful user comments on users' pro-social behavior and the role of their explicit and implicit attitudes within this process. We will pursue this objective using the example of user comments about refugees, because this topic was intensely discussed in the recent public debate during the European refugee crisis (Gagliardone et al., 2015). Accordingly, we will investigate effects on attitudes and pro-social behavior toward refugees, that is, donating money for a refugee relief organization.

1. Attitudinal and behavioral effects of prejudiced messages in user comments

Individuals' attitudes and behavior toward social groups in a society have been shown to be influenced by the depiction of these groups in the media coverage and user-generated content such as user comments online. Research on these effects often relies on Social Identity Theory (SIT) as a theoretical background (Tajfel & Turner, 1986). According to SIT, human beings have a natural tendency of differentiating individuals with whom they (seemingly) share a social category or group membership (ingroup members; e.g., non-migrants) from those whom they perceive to be part of a different group or social category (outgroup; e.g., migrants, refugees). Moreover, humans tend to favor (supposed) ingroup members over outgroup members, i.e., they attribute more positive and more diverse characteristics to ingroup members and less positive and more uniform characteristics to outgroups (Dovidio & Gaertner, 1993; Tajfel & Turner, 1986).

Although this ingroup-outgroup differentiation is thought to be a universal psychological mechanism, the resulting attitudes do not invariably determine actual behavior toward outgroups. Firstly, people can suppress the application of negative attitudes, for example, by acting on other cognitions, such as egalitarian attitudes (Bodenhausen, Macrae, & Sherman, 1999; Devine, 1989; Fiske, 1989). Secondly, negative attitudes must be active to be applied, that is, they need to be triggered. A trigger can be a stimulus that resonates with group-related attitudes in the mental system of an individual (Domke, 2001; Hurwitz & Peffley, 2005). For example, if an individual holds the stereotypic belief that members of an outgroup are often involved in crimes and this person is told a story about a crime, this may automatically make the person think about the outgroup and the negative attributes associated therewith. Even more so, if this person is told that members of the outgroup are frequent criminal offenders, the negative attitude is directly made salient and reinforced.

Media depictions have been shown to be powerful sources of such negative outgroup representations. Stereotypic characterization of, for example, ethnic minorities in mass media or publicly accessible user generated content can activate and reinforce negative attitudes among broad audiences, making these attitudes more likely to influence individual behavior (Dixon, 2008; Domke, 2001; Gilliam & Iyengar, 2000; Power, Murphy, & Coover, 1996). For traditional media content, Gilliam and Iyengar (2000) showed that exposing subjects to a televised crime news story, which identified the suspect as a member of an ethnic outgroup, activated negative attitudes and consequently led to more support for strict punishments. Schemer (2012) discovered a similar pattern studying the tone of media coverage on immigration during a political campaign. More negative coverage was associated with an increase in negative attitudes among the electorate. Such activation of ethnic attitudes in turn can result in more hostile behavior (Amodio & Devine, 2006; Bargh, Chen, & Burrows, 1996; Dijksterhuis & Bargh, 2001). Turning to research on user comments, a similar pattern has emerged in the study from Hsueh et al. (2015). They exposed subjects to antiprejudiced or prejudiced user comments about Asian Americans. The prejudiced comments evoked negative implicit and explicit attitudes towards the target group. In consequence, participants were more likely to engage in negative behavior toward Asian American, that is, they were more likely to write prejudiced comments themselves.

2. Capturing explicit and implicit attitudes

Attitudes that potentially mediate the effect of user comments on behavior can be captured via explicit and implicit attitude measures. Most studies on attitudinal media effects apply explicit measures because such measures are easily implemented, for example, in a questionnaire (Olson, 2009; see Arendt, 2013 for a discussion in communication research). Here, participants are asked to rate explicit statements—e.g., a number of adjectives that describe members of an ethnic group—that reflect their attitude on the object that was rated (Hofmann, Gawronski, Gschwendner, Le, & Schmitt, 2005).

In contrast to explicit evaluations, implicit measures aim at quantifying the closeness of the mental association between the target of an attitude (e.g., members of an ethnic outgroup) and evaluative attributes ("good" vs. "bad"). One of the most prominent examples is the Implicit Association Test (IAT). This test forces subjects to combine the target of the attribute (e.g., African American or White faces or names) with "good" or "bad" labels (Greenwald, McGhee, & Schwartz, 1998). The time it takes to perform the task is used as a measure of the closeness of the mental association between the target and "good" vs. "bad" attributes (i.e., the faster the performance, the closer the mental association, cf. Method section for a detailed account of the procedure).

The investigation of explicit attitudes alone is subject to bias because these attitudes need to be salient to the individual in order to be measurable. Additionally, bias stems from social desirability effects, the latter being especially pronounced when it comes to socially sensitive topics (e.g., ethnic intergroup topics, Greenwald et al., 2009). Furthermore, explicit measures directly suggest the possibility of suppressing or correcting an attitude in its application because they make subjects aware of the attitude's content via explicit propositions (e.g., "refugees are bad"; Arendt, 2013; Gawronski & Bodenhausen, 2006; Greenwald et al., 2002). Implicit attitude measures are, in contrast, less likely to be suppressed or corrected (Arendt, 2013; Greenwald et al., 1998). The reason for that is the fact that implicit attitudes are processed via the activation of associations in the mental system ("refugees", "bad") – a process of which the individual is not necessarily aware (Greenwald et al., 2002). Therefore, studying explicit attitudes only—as it has often been done in research on media effects on ethnic attitudes—may mask important effects stemming from stereotypic messages. As a consequence we seek to analyze the effects of stereotyped messages on both explicit and implicit attitudes:

H1. Negative-civil user comments about refugees result in more negative explicit (H1a) and implicit (H1b) attitudes toward refugees as compared to neutral user comments.

H2. Hateful user comments about refugees result in more negative explicit (H2a) and implicit (H2b) attitudes toward refugees as compared to neutral user comments.

3. Pro-social behavior and ethnic attitudes

Individuals whose negative ethnic attitudes have often been triggered and reinforced will more likely act in line with these cognitions because they are well connected with other parts of the mental system and are therefore easily activated through situational stimuli. Frequently activated attitudes are thus more likely reflected in behavior (e.g., in less helpfulness or support) (Roskos-Ewoldsen, Roskos-Ewoldsen, & Dillmann Carpentier, 2002). Yet even for individuals whose attitudes are rarely activated or who attempt to counter negative attitudes with alternative beliefs and attitudes, it is difficult to entirely suppress the application of a negative attitude when it is directly made salient (e.g., in a user comment) (Blair & Banaji, 1996; Bodenhausen et al., 1999; Devine, 1989; Higgins, 1996).

In the present research, prosocial behavior—namely donating money for refugees—is at the heart of interest. Such prosocial behavior in terms of "actions intended to benefit one or more people other than oneself" (Batson & Powell, 2003, p. 465) has been shown to depend on social group categorization processes: Individuals were more willing to help ingroup as opposed to outgroup members. Yet, re-categorizing outgroup members as belonging to the same superordinate category (e.g., nation) resulted in more prosocial behavior (Gaertner et al., 1999; Penner, Dovidio, Piliavin, & Schroeder, 2005). Beyond these intergroup categorization processes, research suggests that prosocial behavior is sensitive even to entirely subconscious, implicit priming processes (Garcia, Weaver, Moskowitz, & Darley, 2002; Van Baaren, Holland, Kawakami, & Van Knippenberg, 2004). Taken together, one may assume that prosocial behavior differs depending on the attitudes that have been activated within the mental system of the individual—with or without conscious awareness. Based on the arguments and the results of Hsueh et al. (2015) outlined above, we assume negative stereotypes in user comments to activate both explicit and implicit attitudes that, in consequence, decrease prosocial behavior. We assume this to work for both negative (yet civil) statements as well as for outright hate speech.

H3. Negative-civil user comments about refugees indirectly impede prosocial behavior in favor of refugees via explicit (H3a) and implicit (H3b) attitudes.

H4. Hateful user comments about refugees indirectly impede prosocial behavior in favor of refugees via explicit (H4a) and implicit (H4b) attitudes.

4. Differentiating civil negativity and incivility/hate

Negativity in public debates has many faces: While one might criticize that an above-average percentage of refugees would have a below-average educational level and are, in consequence, not able to enter the labor market without difficulties, others might say that refugees would be simply stupid, because they are refugees – which is a rather undifferentiated way to express one's negative attitudes. Uncivil and hateful user comments, of which the latter is an example, have been a particular focal point of the public discourse about the potentially harmful effects of user comments (e.g., Greenblatt, 2015; Lobo, 2015). The term "hateful user comments" here refers to user comments containing severe incivility—i.e., disrespectful speech (e.g., Coe et al., 2014)—directed at individuals with a particular social category affiliation or the social group without any discernible level of differentiation (Gagliardone et al., 2015). In contrast, we use the term "civil negativity" to describe comments that include disagreement and/or opposition in the absence of uncivil language (Chen & Lu, 2017).

Presuming that incivility and hate are particularly harmful is plausible, given that attitudes are more likely activated by obtrusive triggers than by unobtrusive ones (Roskos-Ewoldsen, Klinger, & Roskos-Ewoldsen, 2007). Incivility and hate provoke more attention and emotional arousal than civil negativity alone, since incivility is considered a violation of social norms (e.g., Mutz & Reeves, 2005). Intense cognitive and emotional response states in turn facilitate priming of attitudes through media messages (Chen & Lu, 2017; Valkenburg & Peter, 2013).

Civil negativity, in contrast, could be an "important marker of deliberation" (Chen, 2017; Stromer-Galley, 2007, p. 5), because it illustrates "heterogeneity of perspectives" (ibid.). Still, civil negativity can be sufficient for stimulating negative attitudes and feelings in recipients (Amodio & Devine, 2006). Chen and Lu (2017), for example, showed that both civil and uncivil disagreement in comment sections caused negative emotion and aggressive intentions in readers. The possibility that negative-civil user comments about refugees alone may activate and reinforce negative outgroup attitudes is of crucial importance given that the share of civil negativity in user comments exceeds the share of incivility by far (Ziegele and Quiring, 2017).

In sum, we expect that exposure to both negative-civil and hateful user comments results in negative outgroup attitudes (Chen and Lu, 2017). Still, we expect the effect of hateful comments to exceed the effect of negative-civil comments given that hateful comments are more salient, and constitute stronger appeals than negative-civil appeals alone (for a similar reasoning in the domain of incivility of political actors, see Mutz and Reeves, 2005).

H5. The effect of hateful user comments on attitudes is stronger than the effect of negative-civil comments (H5a: on explicit attitudes; H5b: on implicit attitudes).

5. Method

5.1. Sample and procedure

We conducted a between-subjects online experiment with a single factor that was varied threefold (valence of user comments: hateful, negative-civil, neutral) to test the hypotheses. To estimate the required sample size, we computed an a priori power analysis using the software GPower 3.1 (Faul, Erdfelder, Buchner and Lang, 2009). Based on findings from previous studies on the effects of uncivil and hateful comments (e.g., Anderson, Brossard, Scheufele, Xenos, and Ladwig, 2014; Hsueh et al., 2015; Rösner et al., 2016), we expected small to medium effect sizes of these comments on respondents' explicit/implicit attitudes and on their behavior (max. part. $R^2 = 0.06$). With an $\alpha = 0.05$ and power = 0.95, the projected sample size needed with this effect size was $N = 245$ for between-group comparisons.

Effectively, $N = 253$ participants were recruited via an online access panel. We applied stratified sampling to ensure that the resulting sample was approximately representative of the adult German population regarding gender, age, education (type of school diploma), and region (western vs. eastern German states). Participants mean age was 43.8 years ($SD = 13.7$), 51 percent were female, and 13 percent lived in one of the eastern German states. Thirty-seven percent had graduated from school with a university entrance diploma. Sociodemographic characteristics did not differ significantly between treatment conditions (age: F (2, 250) = 0.36, $p = .70$; gender: χ^2 (2, $N = 253$) = 4.12, $p = .13$; type of school diploma: χ^2 (6, $N = 253$) = 5.52, $p = .48$; region (western vs. eastern German states): χ^2 (2, $N = 253$) = 1.17, $p = .56$).

Participants filled in an online questionnaire and were exposed to mock news articles on the refugee crisis which were accompanied by user comments with fictitious user opinions on that topic. The valence of those user comments was manipulated as part of our experimental design (neutral, negative-civil, hateful). Before reading the news articles and user comments, participants answered basic sociodemographic questions and reported their explicit pre-manipulation attitude toward refugees. After reading the articles and answering buffer questions (evaluation of the news articles), post-manipulation attitudes toward refugees (explicit and implicit) were assessed. Upon completion of the questionnaire, participants received a standard incentive of two Euros for their participation and, as part of the experimental design, an additional five Euros, which they could either donate or keep for themselves (cf. Measures). At the end of the questionnaire, participants were debriefed on the purpose of the study and on the fictitious character of the news articles and user comments.

5.2. Experimental manipulation

In total, three news articles were presented to all participants. The structure and layout corresponded to the presentation of news articles on Facebook. That is, participants were shown the name of the news organization (a fictitious regional newspaper called "Neuer Kurier"), date and time of the posting, a teaser text, and a headline. Beneath, the number of reactions to the articles, the number of users who shared the article, and two user comments were displayed. Two of the three stimulus articles covered the European refugee crisis. One discussed a study about the integration of refugees in the labor market and the spread of right-wing resentments in the German population. The other was about a man associated with a radical nationalist organization who physically assaulted a refugee in a German city. Between those two articles on the European refugee crisis, we showed participants an article that covered an accident at a railroad crossing. Showing an article which was not concerned with the refugee topic served the purpose of diverting participants' attention.

The first of the user comments posted at the bottom of the two articles about the European refugee crisis served as the experimental manipulation. Participants were randomly assigned to comments that either addressed refugees in a civil and neutral manner (condition: neutral user comment, e.g., "it's up to us whether we want to allow refugees to make a good life for themselves here"), in a negative, but civil manner (condition: negative-civil user comment, e.g., "unfortunately, many of them [refugees] are quite aggressive as well and often their primary motive is financial benefit"), or in an uncivil, derogative, and hateful manner (condition: hateful user comment, e.g. "this raunchy mob [refugees] is endangering everyone … aggressive as hell, bloody stupid, and only interested in ripping us off!"). The second comment posted to each of the stories about the refugee crisis and both comments posted under the buffer story were neutral in tone and identical across the experimental conditions.

5.3. Measures

Post-manipulation implicit attitudes toward refugees. Implicit attitudes toward refugees were gauged using the IAT (Greenwald et al., 1998) adapted to refugees (for more details, see the appendix). The IAT measures the relative strength of the association between two target concepts (refugees vs. Germans) and pleasant ("good") vs. unpleasant ("bad") attributes. The underlying assumption is that when participants are forced to perform an action that connects a concept (e.g., refugees) with an attribute of a specific valence (e.g., "good") they will be faster in doing so if the concept is consistent with this attribute in participants' mental representation of social reality (i.e., when they indeed have a favorable attitude toward refugees).

Following the standard procedure (Greenwald et al., 1998), the IAT started with three task blocks which were introduced as "training"--blocks (the first two being actually used for training purposes, the third being relevant for the analysis). In the first block, the IAT displayed pictures representing the two target concepts (pictures of apparent refugees or of apparent Germans, one picture at a time, 10 trials). On the left side of the screen, participants saw a label for one of the target concepts (the word "German"). On the right side they saw the label for the other concept (the word "refugee"). Participants were instructed to hit a key on the left side of the keyboard (the "e"-key) whenever they were shown a picture of someone who appeared to be German. Whenever they saw a picture of someone appearing to be a refugee they were supposed to hit a key on the right side of the keyboard (the "i"-key). The time it took participants to hit one of the keys was recorded for each trial (in milliseconds: ms). In the second block, participants were presented words at the center of the computer screen that were either unambiguously "good" or "bad" (10 trials). On the left side of the screen participants saw the label "good" and on the right side the label "bad". Whenever participants were shown a word they were supposed to hit the "e"-key for "good" words or the "i"-key for "bad" ones.

The third block required participants to look at either a picture (an apparent German or an apparent refugee) or a word (a "good" or a "bad" word, 10 trials). At the left side of the screen, the labels "German" and "good" were simultaneously displayed. On the right side, participants saw the labels "refugee" and "bad". Whenever they were shown a picture of someone appearing to be a German or a "good" word they were supposed to hit the "e"-key. For pictures of apparent refugees or "bad" words they should hit the "i"-key. After that, a forth block was introduced as the actual test task with the same structure as the third block, but with 30 trials.

The fifth and sixth task blocks were again introduced as training tasks (with the sixth block actually being relevant for the analysis). In the fifth block, only pictures of apparent refugees and apparent Germans were displayed (10 trials) but the placing of the labels was reversed ("refugees" on the left, connected to the "e"-key, "German" on the right, connected to the "i"-key). During the sixth block, participants were shown pictures or words (10 trials) with the labels "refugee" and "good" displayed on the left side of the screen ("e"-key) and the labels "German" and "bad" on the right side ("i"-key). The final block was equivalent to the sixth but it was introduced to participants as the second part of the actual test and comprised 30 trials.

The IAT test scores were calculated following the procedure proposed by Greenwald, Nosek, and Banaji (2003), the improved scoring algorithm for the d score, based on the response latencies produced in task blocks three, four, six, and seven. First, we checked that no more than 10 percent of the responses were given within less than 300 ms (if that were the case, the test would not be valid). The 10th percentile was no lower than 600 ms for any of the trials in blocks three, four, six, and seven. Afterwards we deleted response latencies of more than 10 s and calculated standard deviations jointly for all response latencies in blocks three and sixth and in blocks four and seven as well as mean response latencies for each of the four blocks separately. Then, we identified false answers (wrong key selected). These false answers were replaced by the mean block response time plus a penalty of 600 ms. After the penalties were introduced, we produced new mean response latencies (including the penalties). Then we calculated the differences between the new mean response latencies of block six minus block three and block seven minus block four. These two scores were then standardized using the joint standard deviations (blocks 3 and 6, blocks 4 and 7). As a last step, the mean value of the two standardized scores was calculated as the overall IAT score and multiplied by -1. In the resulting measure, smaller values express larger favoritism of Germans over refugees and larger values express less favoritism. In other words, larger values represent more positive implicit attitudes toward refugees.

Post-manipulation explicit attitudes toward refugees. Explicit attitudes were established by asking participants (after stimulus exposure) to rate eight pairs of adjectives. For each pair, participants indicated on 7-point scales which adjective better described refugees (e.g., *1*: 'ungrateful' – *7*: 'grateful'; *1*: 'stupid' – *7*: 'intelligent'; $\alpha = 0.92$; Bargh and Pietromonaco, 1982; Srull and Wyer, 1980). The scale will be established via confirmatory factor analysis.

Pre-manipulation explicit attitudes toward refugees. To control for participants' attitudes prior to exposure to the news articles and user comments we asked them (before stimulus exposure) to rate five statements representing explicit resentment toward refugees (e.g., "refugees are a threat to society", "refugees should receive the same social benefits as Germans") on seven-point rating scales (*1*: 'does not apply at all' – *7*: 'fully applies'; $\alpha = 0.90$; based on Schemer, 2012; Gilliam and Iyengar, 2000). We used different sets of indicators for measuring pre- and post-manipulation attitudes to avoid bias from consistency effects in participants' post-exposure answers. The scale will be established via confirmatory factor analysis.

Pro-social behavior. In addition to their standard incentive of two Euros, participants received five additional Euros. Participants were asked whether they wanted to donate these five Euros, or a share of it, to a humanitarian non-governmental organization dedicated to helping refugees (a German partner organization of the UNHCR: Deutsche UNO-Flüchtlingshilfe). Participants decided on a sliding scale how many cents of the five Euros they wanted to donate/to be payed as an additional incentive (between 0 and 500 Cents in 1 Cent units).

Treatment check. Participants were presented a list of adjectives (factual, aggressive, hateful, derogatory, positive). They were asked to what extent they agreed that these adjectives described the user comments they had read (seven-point rating scale; *1*: 'I do not agree at all' – *7*: 'I fully agree'). Using the same response scale, participants were asked to indicate their agreement with the statement 'such user comments are likely to be found under news articles from other news outlets as well' and whether the comments violated norms and values. Finally, participants reported whether the user comments had changed their emotional state.

Sociodemographic characteristics. We asked participants to indicate their gender, age, the type of school diploma they held (if any), and their home region.

Means and standard deviations for all variables are shown in Table 1.

Table 1
Means and standard deviations.

	Mean (standard deviation)
Treatment check	
Factual	3.6 (1.8)
Aggressive	3.9 (2.1)
Hateful	3.9 (2.1)
Derogatory	3.7 (2.1)
Positive	3.3 (1.7)
Norm-violating	5.3 (1.5)
Mood-changing	4.0 (2.1)
Evaluation of comments as realistic	5.3 (1.5)
Control variable	
Pre-manipulation explicit attitudes (scale)[a]	2.6 (1.7)
Refugees drain the welfare state	3.9 (2.0)
With all those refugees I feel like a stranger in my own country	3.9 (2.3)
Refugees should receive the same social benefits as Germans	2.9 (1.8)
Refugees are a threat to society	3.6 (2.0)
Germany should receive less refugees	4.9 (2.0)
Mediators and dependent variables	
Post-manipulation explicit attitude (scale)[a]	2.8 (1.0)
Refugees are:	
Likeable vs. unlikeable	4.1 (1.3)
Lazy vs. hard working	4.2 (1.3)
Ungrateful vs. grateful	4.2 (1.6)
Impossible to integrate vs. easily integrated	4.0 (1.7)
Honest vs. dishonest	4.1 (1.3)
Aggressive vs. peaceful	4.0 (1.4)
Stupid vs. intelligent	4.2 (1.2)
Uneducated vs. educated	3.9 (1.3)
Post-manipulation implicit attitudes (IAT)	0.5485 (0.4848)
Prosocial behavior	100.6 (158.8)

[a] Scores obtained via regression imputation following confirmatory factor analysis.

6. Results

6.1. Treatment check

As preliminary analysis, we checked whether participants had noticed the valence of the user comments they had read. Indeed, participants in the three treatment conditions rated the comments significantly different regarding all proposed adjectives (factual: F (2, 250) = 26.96, $p < .01$; aggressive: F (2, 250) = 105.60, $p < .01$; hateful: F (2, 250) = 100.71, $p < .01$; derogatory: F (2, 250) = 94.43, $p < .01$; positive: F (2, 250) = 46.67, $p < .01$). In all five cases, Bonferroni post-hoc tests revealed that negative-civil user comments were perceived as significantly less factual and positive, but as more aggressive, hateful, and derogatory than the neutral comments. The hateful comments were perceived as less factual and positive, but as more aggressive, hateful, and derogatory than both the neutral and the negative-civil user comments.

Furthermore, participants in the three treatment conditions differed regarding the level of norm violation ascribed to the user comments (F (2, 250) = 28.67, $p < .01$). The hateful comments were rated, on average, as norm violations ($M = 5.26$, $SD = 1.80$) compared to neutral comments ($M = 3.16$, $SD = 1.77$) or negative-civil comments ($M = 4.16$, $SD = 1.78$). According to Bonferroni post-hoc tests, the rating of the negative-civil comments differed significantly from the rating of the neutral comments. The hateful comments were perceived as significantly more norm-violating than the neutral and the negative-civil comments. The same pattern of results emerged regarding participants' mood. The treatment conditions differed significantly (F (2, 250) = 22.41, $p < .01$) with participants exposed to neutral comments negating a change in their mood ($M = 3.03$, $SD = 1.62$) and participants exposed to hateful comments affirming it ($M = 4.99$, $SD = 2.17$). Participants in the negative-civil condition were, on average, undecided ($M = 4.22$, $SD = 1.97$). Post-hoc tests again showed that the negative-civil condition differed significantly from the neutral condition. The hateful condition differed significantly from both its negative-civil and neutral counterparts.

Despite these differences, the three types of comments were perceived as equally realistic. Participants' agreement with the statement, “such user comments are likely to be found under news articles from other news outlets as well”, did not differ significantly between treatment conditions (F (2, 250) = 0.56, p = .57). Average scores above the center scale indicate that neutral comments (M = 5.14, SD = 1.55), negative-civil comments (M = 5.24, SD = 1.51), and hateful comments (M = 5.39, SD = 1.51) were perceived as realistic.

6.2. Confirmatory factor analysis: explicit attitudes pre- and post-manipulation

To confirm the assumed structure of the measures representing pre-manipulation (five items, one factor) and post-manipulation explicit attitudes toward refugees (eight items, one factor), we conducted a confirmatory factor analysis (CFA) with AMOS 23. Pre- and post-manipulation explicit attitudes were modelled as one latent variable each with the designated items loading on these latent variables. Implicit attitudes were added to the model to enable assessment of discriminant validity. Overall model fit was satisfactory with a comparative fit index (*CFI*) of 0.96, a root mean squared error of approximation (*RMSEA*) of 0.05, a standardized root mean squared residual (*SRMR*) of 0.05. However, the χ^2/df-ratio was slightly above the desired threshold of 2 with $\chi^2/\text{df} = 2.18, p < .01$ (procedure for assessing overall model fit: Hu and Bentler, 1999). Modification indices revealed that the error terms of two indicators (stupid vs. intelligent; uneducated vs. educated) for post-exposure explicit attitudes were correlated. The correlation is conceptually plausible as both indicators refer to the competence of refugees whereas the other six indicators assessed character traits that reflect a warmth dimension (e.g., ungrateful vs. grateful; honest vs. dishonest; Cuddy, Fiske and Glick, 2008). We therefore added the correlation to the CFA. This adjustment resulted in the overall fit of the model: χ^2 (148) = 263.88 ($\chi^2/df = 1.78$, $p < .01$); $CFI = 0.98$; $RMSEA = 0.04$; $SRMR = 0.04$). Squared multiple correlations (*SMC*), as indicators for individual item reliability, exceeded 0.50 (i.e., the factor explained more than 50% of the item’s variance) for all but three items. One item came close to 0.50 with an *SMC* of 0.49. The two remaining items (one in each factor) still had an *SMC* of above 0.30. Reliability of the factors was assessed drawing on their composite reliability (*CR*) and average variance extracted (*AVE*). Post-manipulation explicit attitudes had a *CR* of 0.92, for pre-manipulation explicit attitudes *CR* was 0.91 (considerably above the suggested minimum *CR* of 0.60). *AVE* was 0.59 for post- and 0.66 for pre-manipulation attitudes (above the suggested minimum of 0.50). Discriminant validity was satisfactory with both *AVE* values exceeding the squared correlations between the three attitude measures (criteria applied as suggested in Bagozzi and Yi, 1988 and Fornell and Larcker, 1981, cf. Table 2 and Table 3).

Table 2
Confirmatory factor analysis.

	CR	AVE	Factor loading	SMC
Pre-manipulation explicit attitudes (scale)[1]	.91	.66		
Refugees drain the welfare state			.87	.75
With all those refugees I feel like a stranger in my own country			.88	.77
Refugees should receive the same social benefits as Germans			-.58	.34
Refugees are a threat to society			.89	.78
Germany should receive less refugees			.82	.67
Post-manipulation explicit attitude (scale)	.91	.59		
Refugees are:				
Likeable vs. unlikeable			.83	.69
Lazy vs. hard working			.77	.59
Ungrateful vs. grateful			.85	.72
Impossible to integrate vs. easily integrated			.80	.64
Honest vs. dishonest			-.61	.37
Aggressive vs. peaceful			.82	.67
Stupid vs. intelligent			.70	.49
Uneducated vs. educated			.73	.53

CR = Composite Reliability; AVE = Average Variance Extracted; SMC: Squared Multiple Correlations.

Table 3
Zero-order correlations.

	1	2	3	4	5	6
1: Negative-civil comments	1	-.45*	.01	-.08	.07	-.07
2: Hateful comments		1	.03	-.05	-.17*	-.09
3: Pre-manipulation explicit attitudes			1	-.71*	-.21*	-.39*
4: Post-manipulation explicit attitudes				1	.14*	.30*
5: Post-manipulation implicit attitudes (IAT)					1	.20*
6: Prosocial behavior						1

*p < .05.

Table 4
Mean differences.

Tone of user comments	Post-manipulation explicit attitudes	Post-manipulation implicit attitudes	Prosocial behavior (cents donated)
Neutral	.20^{a}	.11^{a}	131
Negative-civil	-.16^{b}	.10^{a,b}	86
Hateful	-.06a,^{b}	-.26^{b}	78
η^2	.03	.03	.02
p	.04	.03	.06

ANOVA; Different superscripts represent statistically significant differences with $p < .05$. Implicit and explicit post-manipulation attitudes are controlled for pre-manipulation explicit attitudes (standardized residuals).

6.3. Description of relevant mean differences

Prior to testing the mediation model proposed in our hypotheses, we will present the mean differences for three focal comparisons. Our hypotheses assume that civil negativity and hate in user comments about refugees should alter (a) explicit and (b) implicit attitudes toward refugees as well as (c) the level of prosocial behavior in favor of refugees. We therefore compared participants confronted with hateful, negative-civil, or neutral user comments regarding the amount of money they donated to a refugee relief organization (prosocial behavior) and regarding their explicit and implicit attitudes toward refugees after exposure to the user comments (post-manipulation) while controlling for the pre-manipulation level of the explicit attitudes (ANOVA, Table 4, Figs. 1–3).

Results show that the tone of the user comments significantly affected post-manipulation explicit and implicit attitudes. Specifically, hateful comments resulted in more negative implicit attitudes toward refugees than neutral comments. Negative-civil user comments resulted in more negative explicit attitudes toward refugees than neutral user comments. In total, the tone of the user comments explained about three percent in the variance of post-manipulation implicit and explicit attitudes (η^2).

In this simple comparison, the effect of the tone of user comments on prosocial behavior (amount of money donated) fell short of significance (p = .06), with the most money being donated by participants confronted with neutral user comments (M = 131 cents; SD = 183), followed by participants who read the negative-civil comments (M = 86 cents; SD = 146) or the hateful comments (M = 78 cents; SD = 133).

6.4. Testing hypotheses

To test our hypotheses, we calculated a structural equation model (SEM) with AMOS 23. Zero-order correlations for all variables included in the model are displayed in Table 3. Tone of user comments was entered via two dummy-coded variables (negative-civil tone = 1; hateful

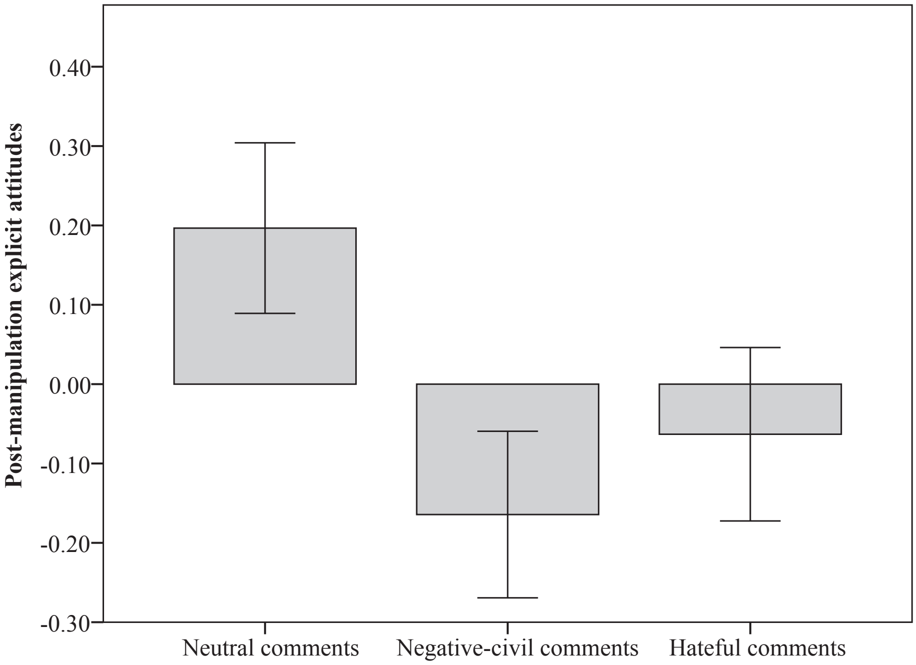


Fig. 1. Mean differences; post-manipulation attitudes controlled for pre-manipulation attitudes (standardized residuals); error bars represent standard errors.

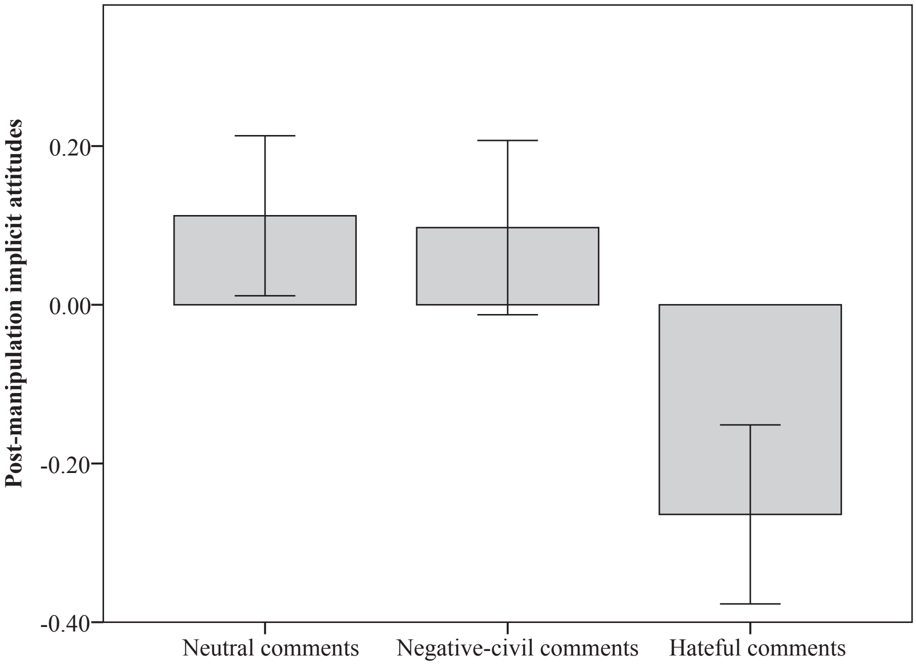


Fig. 2. Mean differences; post-manipulation attitudes controlled for pre-manipulation attitudes (standardized residuals); error bars represent standard errors.

tone = 1; neutral tone: reference category for both dummy variables). These two dummy variables served as predictors for the post-manipulation attitudes (implicit and explicit). The post-manipulation-attitudes in turn were modelled as predictors for participants' pro-social behavior (amount of money donated). Pre-manipulation attitudes were entered as a control (predicting both implicit and explicit post-manipulation attitudes). To test a full mediation model, direct effects of both types of user comments on pro-social behavior were also estimated. For the purpose of parsimony, the control variable (pre-manipulation explicit attitudes toward refugees) is not shown in Fig. 4 and latent variables are depicted without their indicators. Significance of indirect effects was established by bootstrapping of bias-corrected confidence intervals (5000 samples).

The model fit the data well (χ^2 (110) = 158.43 (χ^2/df = 1.44), $p < .01$; $CFI = 0.98$; $RMSEA = 0.04$; $SRMR = 0.05$). Exposure to negative-civil tone in user comments resulted in significantly more negative explicit attitudes toward refugees as compared to exposure to neutral comments ($\beta = -0.11$, $p = .04$; H1a supported). However, negative-civil comments (relative to neutral comments) did not affect implicit attitudes ($\beta = -0.01$, $p = .91$; H1b rejected). Both explicit ($\beta = 0.28$, $p < .01$) and implicit attitudes ($\beta = 0.15$, $p = .01$) were significantly related to pro-social behavior. Accordingly, the negative-civil comments indirectly impaired pro-social behavior via participants' explicit attitudes ($\beta = -0.03$, $p = .04$ H3a supported), yet not via

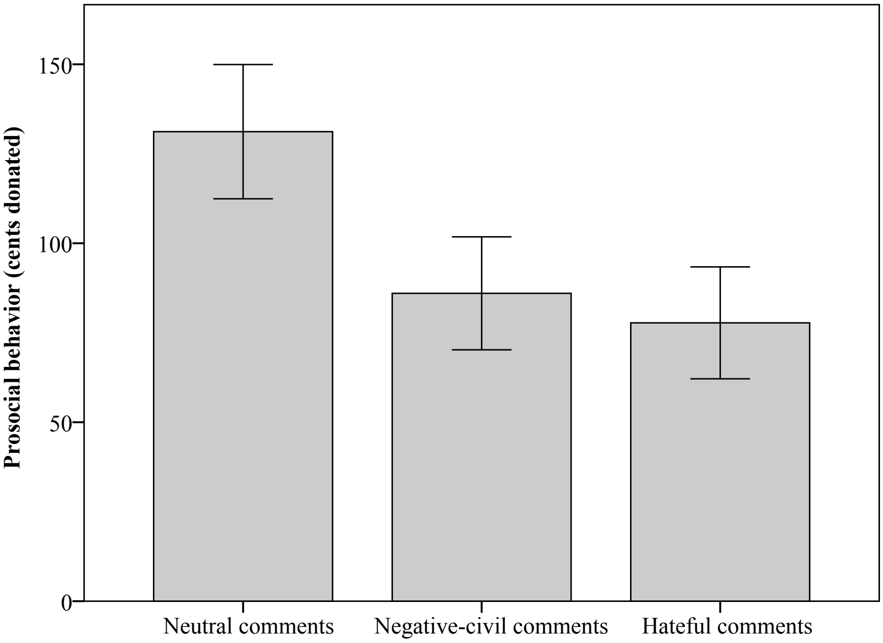


Fig. 3. Mean differences; error bars represent standard errors.

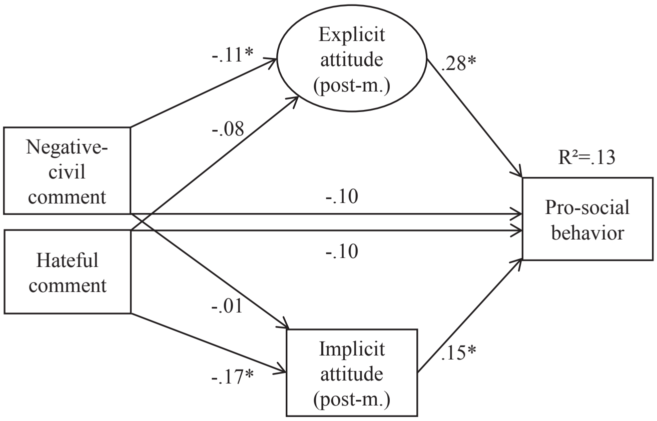


Fig. 4. Structural equation model; *p < .05; standardized regression weights; regression weights for the control variable (pre-manipulation explicit attitude) on post-manipulation explicit attitudes: 0.71, p < .05; on post-manipulation implicit attitudes: 0.20, p < .05; model fit: χ^2 (110) = 158.43 (χ^2/df = 1.44), p < .01; CFI = 0.98; $RMSEA$ = 0.04; $SRMR$ = 0.05.

implicit attitudes ($\beta < 0.01$ p = .86; H3b rejected). The direct effect of negative-civil comments on prosocial behavior was not significant, either ($\beta = -0.10$, p = .14). Still, negative civil-comments exerted a total effect on pro-social behavior that only just fell short of significance ($\beta = -0.13$, p = .06).

The effects of hateful user comments showed a reversed pattern. A hateful tone was unrelated to explicit attitudes toward refugees ($\beta = -0.08$, p = .17; H2a rejected), yet it significantly increased negative implicit attitudes ($\beta = -0.17$, p = .01; H2b supported). Hateful comments, hence, had a significant indirect effect on pro-social behavior via implicit attitudes ($\beta = -0.03$, p = .01; H4b supported), but not via explicit attitudes ($\beta = -0.02$, p = .14; H4a rejected). The direct effect of hateful user comments on prosocial behavior failed significance ($\beta = -0.10$, p = .15). In sum, hateful user comments had a significant total effect on prosocial behavior ($\beta = -0.14$, p = .02). Altogether, the model explains thirteen percent of the variance in our central dependent variable (prosocial behavior).

In hypothesis five, we assumed that hateful comments should have a stronger influence on both explicit (H5a) and implicit attitudes toward refugees (H5b) than negative-civil comments. Two alternative models were calculated to test these hypotheses. In the first model, we imposed an equality constraint on the path coefficients corresponding to H1a (negative-civil comment → explicit attitude) and H2a (hateful comment → explicit attitude). Comparison with the model without constraints revealed that the two models did not differ significantly (χ_{diff} (1) = 0.25, p = .62, H5a rejected). This means that the effect of negative-civil user comments on explicit attitudes was not significantly different from the effect of hateful user comments on explicit attitudes. In the second model, we defined the path coefficients corresponding to H1b (negative civil-comment → implicit attitude) and H2b (hateful civil-comment → implicit attitude) to be equal. Model comparison showed that this model differed significantly from the model without constraint (χ^2_{diff}(1) = 5.22, p = .02). Hence the two paths coefficients significantly differed from one another (standardized path coefficient for hateful comments, $\beta = -0.17$, and for negative-civil comments, $\beta = -0.01$). Thus, hateful comments had a stronger impact on implicit attitudes than negative-civil comments, supporting H5b.

7. Discussion

Incivility in user comments purportedly poses a threat to democratic societies. Particularly hate speech as an extreme form of incivility is regarded as detrimental for the social cohesion between different strata of pluralistic societies (e.g., Anderson, Scheufele, Xenos, and Ladwig, 2014; Greenblatt, 2015; Lobo, 2015). It was the aim of the current research to add substantial empirical evidence to this debate by studying the attitudinal and behavioral effects of negative-civil and uncivil/hateful user comments on social minorities. In doing so, the present study relied on the behavioral measure of donating money for an aid organization as its dependent variable which represents a real and consequential type of behavior. Additionally, it was the aim to investigate whether the effects arise from incivility in terms of hateful comments only, or whether negative, yet civil comments may result in comparable effects on recipients' pro-social behavior. Finally, attitudes toward refugees were included as mediators – differentiating between explicit and implicit attitudes.

The results reveal that both negative-civil and hateful user comments against refugees negatively impacted participants' attitudes and, in turn, also impeded their pro-social behavior toward a refugee relief organization. Therefore, one can conclude that reading dismissive user comments against members of an outgroup can be harmful for intergroup behavior because negative intergroup attitudes are activated (in line with the results of Hsueh et al., 2015). Although both negative-civil and hateful user comments affected recipients' attitudes and their prosocial behavior, the effects were more differentiated: While negative-civil user comments had a significant effect on explicit attitudes, hateful user comments impacted implicit negative attitudes toward refugees – with both exerting indirect effects on prosocial behavior. Additionally, implicit attitudes were more strongly affected by hateful as compared to negative-civil user comments.

Due to the universal psychological mechanism of ingroup-outgroup differentiation, stereotyped messages have the potential to activate and reinforce negative attitudes, which in turn condense in corresponding behavioral responses (as described by Domke, 2001; Hurwitz and Peffley, 2005). Prosocial behavior in particular has been shown to be influenced by ethnic attitudes and seems to be sensitive to both implicit and explicit attitudes. These effects resulted from both negative-civil and hateful user comments which let us conclude that user comments may indeed substantially influence the social fabric in a society (Hsueh et al., 2015; Winter, Brückner and Krämer, 2015). As a consequence, public concerns on the potential threat of user comments for democratic societies seems to fall short, because they focus on incivility and outright hate speech only (e.g., Greenblatt, 2015; Lobo, 2015). Our results indicate that even civil criticism affects recipients' prosocial behavior, which may result in less social cohesion and intergroup solidarity. This is even more troubling since negative, but civil statements outnumber uncivil user comments by far (Ziegele and Quiring, 2017) and are not deleted from news outlets' websites or Facebook pages as it is often the case for hateful comments (Domingo, 2011).

While a critical public stance on current political developments such as immigration policies is regarded as important for the public discourse in terms of democratic deliberation, it seems like the same critique may push intergroup polarization processes. In order to avoid such detrimental social effects, measures have to be developed to counter user comments that activate negative attitudes via mere criticism or even hate speech. One possibility of doing so may be to engage in counter arguing or counter speech, respectively (Braddock, 2016; Macnair and Frank, 2017). This can be done via more active community management and active moderation of user discussions below journalistic articles (e.g., Stroud et al., 2015; Ziegele and Jost, 2016). Counter arguing may also be accomplished by the community itself with courageous and dedicated users posting messages that present counter narratives to the stereotypic messages (Ziegele, Naab, & Jost, 2019).

Two main conclusions can be drawn from the differential effects of negative-civil and hateful user comments on explicit and implicit attitudes: First, implicit and explicit attitudes do not seem to be just two sides of the same coin; they are not interchangeable (Arendt, 2013; Arendt, Marquart and Matthes, 2015). They were differentially sensitive to different stimuli with both significantly predicting the same type of behavior. Hence, to get the full story, one has to rely on both concepts—the activation of implicit associations towards an outgroup and the activation of the explicit manifestation of an attitude.

Second, negative-civil and hateful messages indirectly resulted in the same behavioral response (i.e. reduced prosocial behavior), yet they seemed to be channeled via different cognitive processes. Even more so: implicit attitudes were significantly more strongly affected by hateful as compared to negative-civil attitudes. This result is in line with the implications from priming research according to which a more obtrusive priming stimulus, arousing more attention and emotional responses, results in more pronounced attitudinal effects as compared to an unobtrusive stimulus (here: the negative-civil user comments) (Roskos-Ewoldsen et al., 2007). Still, the difference in effects between negative-civil and hateful comments on explicit attitudes falls short of significance. One reason might be the fact that explicit attitudes are based on more conscious processing and elaboration. Hence, cognitive corrective actions might have been at play here (Arendt et al., 2015). While those corrections were applied to hateful comments, negative-civil comments might not have activated the need to correct the message as much, because they did not arouse enough attention and are less subject to social sanctioning (Roskos-Ewoldsen et al., 2007). This also suggests that the effect of hateful comments on implicit attitudes, however, seems to result from automatic processing of messages (Gawronski and Bodenhausen, 2006). Particularly these unconscious effects should be seen with concern: While conscious reasoning can be subject to active corrections by the individual (i.e., active negation, active counter arguing while elaborating), the seemingly unconscious effects on implicit attitudes are beyond the scope of such corrective cognitive actions (Gawronski and Bodenhausen, 2006). In strengthening recipients' media literacy (Arendt, 2013), one may enforce such corrective cognitive actions rendering the effects of civil criticism less consequential for recipients' prosocial behavior. Yet, with respect to hate in user comments, such conscious cognitive counter strategies will supposedly be only of limited success: Comparable to the results of Arendt et al. (2015), recipients are affected with respect to their implicit attitudes and their prosocial behavior and they even might not know or realize what is happening to them.

Nevertheless, our results have to be interpreted in light of some limitations: Our experimental design allows for studying short-term effects of user comments on recipients' attitudes and their prosocial behavior. In the long run, such effects may lose considerably in strength (as has been shown for priming effects, Roskos-Ewoldsen et al., 2007) or even change in light of exposure to more nuanced and diverse stimuli. Nevertheless, hateful comments and particularly negative-civil comments seem to be widespread in online user discussions (Coe et al., 2014; Rowe, 2015; Su et al., 2018). Therefore, one may assume that recipients are exposed to comparable messages on a regular basis. Particularly, the current spread of right-wing populist voices in the public debates of Western societies is worrisome in this regard: Since the rejection of social minorities and from time to time even outright hatred are part of a constant stream of publicly spread messages by these political actors via digital platforms such as Twitter or Facebook, the effects detected in our study might be potentiated due to the role model character (at least in some parts of the society) of prominent populist actors, such as Donald Trump in the US, Matteo Salvini in Italy or Heinz-Christian Strache in Austria. Consequently, research designs should be employed which are able to detect such cumulative and long term effects of negativity and hate speech on digital communication platforms in general and in user comments in particular.

Second, we have obtained our results with respect to user comments on the European refugee crisis. One might argue that the effects of user comments on attitudes might have been less pronounced and less consequential with respect to actual behavior for a topic which was not accompanied by so much involvement and polarized public opinion as it was the case for the refugee topic (e.g., Harteveld, Schaper, De Lange and Van Der Brug, 2018).

Third, in terms of measurement of implicit attitudes our aim of research required the construction of a specific type of IAT, in which refugees as a specific type of outgroup were visually fronted with ingroup members of the German society. Although the analysis of our participants' test results (correctly identifying Germans vs. refugees on the pictures) and the results of our validation study (see appendix) indicated the robustness of our findings, the question remains open of how to construct a valid and more general measure for implicit attitudes towards different (out)groups in a society. We have chosen this topic in order to present our participants a realistic scenario and hence, ensuring a high external validity of results. Even more so: We have obtained our results for a sample, which is more diverse as compared to the student or convenience samples usually analyzed in social sciences. The participants in this study match the distribution criteria regarding gender, age, education, and home region of the German population. Therefore, our results can be regarded as substantial and high in external validity. Still, there is a need for more evidence on a broader range of topics.

In conclusion, our results suggest that negative stereotyped user comments, which are widespread in today's interactive online environment, have worrisome consequences on attitudes and real, consequential pro-social behavior. Particularly the fact that different psychological channels are involved in transmitting the effects of negative and hateful user comments poses a challenge for those who aim at taming these negative effects of user comments on other recipients. Simply banning specific content from online discussion sections (e.g., hate speech against specific social groups) does not seem to be sensitive enough to thwart all the nuanced detrimental social effects which may arise from negatively stereotyped messages. Even more so, banning may curb free speech. Journalists, community managers, politicians, and scientists are, therefore, asked to develop ideas and measures, which are more nuanced to the sources of such harmful effects to enable more constructive discussions among users in online environments. This could be a way to realize some of those positive promises of the internet for democratic societies instead of constantly trying to eradicate all those negative by-products of online environments.

Acknowledgments

This work was supported by the Center for Media Convergence at the Johannes Gutenberg University of Mainz.

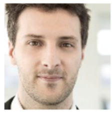
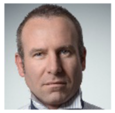
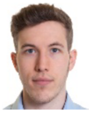
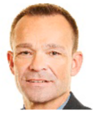
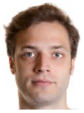
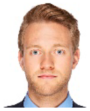
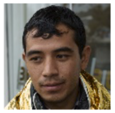
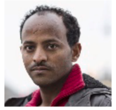
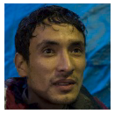
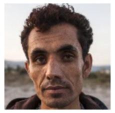
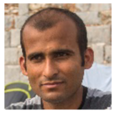
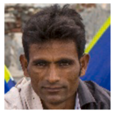
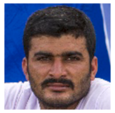
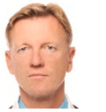
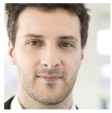
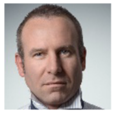
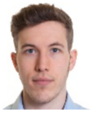
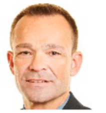
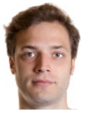
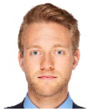
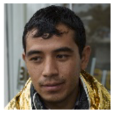
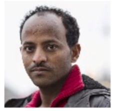
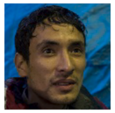
Appendix

The IAT that was applied is a mixture of the classic verbal IAT and a pictorial IAT. From the verbal IAT, we borrowed the task that requires the participants to categorize positive (magnificent, amazing, fabulous, outstanding, fantastic, awesome, terrific) and negative words (horrible, disgusting, repulsive, terrible, dreadful, appalling, upsetting) as positive or negative (Olson & Fazio, 2006). Since it was difficult to find words that can be used for the task to identify refugees as a group, we relied on a task that is often used in pictorial IATs to assess social group stereotypes. For instance, in the native American IAT, participants are shown pictures of European and Indian American individuals and are required to categorize the pictures as American or foreign (Nosek et al., 2007). The Asian American IAT works similar.

Transferred to refugees, we had our participants look at pictures that show refugees (outgroup: the pictures contain symbols that are characteristic for the visual representation of refugees during the migration crisis; e.g. thermal blankets, flysheet) with darker skin tone (based on the ethnic composition of the refugees coming to Germany in 2015: 67% Arab, 25% Kurdish; BAMF, 2016) and Germans with light skin tone (ingroup) and categorize the pictures as German or refugee. Since we wanted to assess implicit attitudes and not stereotypes, we combined the pictorial categorization task and the evaluative word categorization task.

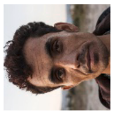
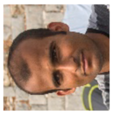
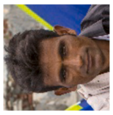
We also conducted a post-hoc validation study to show that the pictures used are similar in perceived valence. Sixty participants (80% female, Age $M = 21.53$, $SD = 4.66$) looked at seven pictures of refugees and seven pictures of Germans and rated the individuals on two semantic differentials, i.e., dislikable – likable, negative – positive (seven-point rating scales). The items were summarized per picture to form a composite score of perceived likability (Cronbach's α ranging between 0.72 and 0.90 for all pairs of ratings). A repeated measures analysis of variance demonstrates that there are differences between picture ratings, F (13, 754) = 35.66, $p < .01$. Table A1 exhibits Bonferroni-corrected mean difference tests across all picture evaluations. Specifically, the first, the third and the fifth picture of ingroup members were perceived more negatively. Similarly, the last outgroup picture was rated more dislikable than all other outgroup members. For the calculation of the IAT score, we used the reaction times to all trials in blocks 3, 4, 6 and 7, although the likability ratings partly differed. The differences in perceived valence make it more difficult to falsify our hypothesis since the IAT includes more likable outgroup members than ingroup members. In order to show that the inclusion of all pictures does not invalidate the IAT, we conducted all tests with an alternative calculation of the IAT score. This alternative d score was calculated on the basis of reaction times of pictures that do not differ regarding their valence (pictures 2, 4, 6 for the ingroup, and pictures 9, 11, 12 for the outgroup). We could have used more reaction times for outgroup pictures, but this would have meant to use an unequal number of outgroup and ingroup pictures for the calculation of the IAT score. Using this more restrictive IAT score did not change the substantive findings. Specifically, exposure to hateful comments increased negative implicit attitudes (β = −.12, p <. 01), but exposure to negative-civil comments affected negative implicit attitudes only marginally (β = −0.09, $p = .70$). The direct effect of implicit attitudes on donation behavior was also similar in size (β = 0.12, $p < .01$) indicating that higher implicit ingroup favoritism reduced the amount of money donated to the refugee relief organization. These robustness checks lent credence to the validity of the IAT and the related findings. The data can be obtained from the authors upon request.

Table A1
Mean Difference in Likability Ratings Across Pictures of Ingroup (pictures 1–7 in the first column) and Outgroup Members (pictures 8–14 in the first column)

	Ingroup 2	Ingroup 3	Ingroup 4	Ingroup 5	Ingroup 6	Ingroup 7	Outgroup 1	Outgroup 2	Outgroup 3	Outgroup 4	Outgroup 5	Outgroup 6	Outgroup 7
Ingroup 1													
	1.50*	.72*	−1.37*	-.73	−1.65*	−1.61*-	−1.61*	−1.31*	−1.50*	−1.22*	−1.24*	-.98*	-.35
		2.22*	.13	.76*	-.15	-.11	-.11	1.9	.01	.28	.26	.53	1.15*
			−2.09*	−1.46*	−2.73*	−2.33*	−2.03*	−2.33*	−2.22*	−1.94*	−1.96*	−1.69*	−1.07*
				.64	-.28	-.24	-.06	-.24	-.13	.15	.14	.39	1.03*
					-.92*	-.87*	-.84*	-.58	-.76*	-.48	-.50	-.24	.39
						.04	.04	.34	.15	.43	.42	.68*	1.30*
							.01	.29	.11	.39	.37	.64*	1.26*
								.30	.10	.38	.39	.60*	1.60*
									-.19	.09	.08	.40	.97*
										.28	.26	.53	1.12*
											-.02	.25	.87*

(continued on next page)

Table A1 (*continued*)

	Ingroup 2	Ingroup 3	Ingroup 4	Ingroup 5	Ingroup 6	Ingroup 7	Outgroup 1	Outgroup 2	Outgroup 3	Outgroup 4	Outgroup 5	Outgroup 6	Outgroup 7
													
												.26	.89*
													-.63*

Note: Statistically significant Bonferroni-corrected mean differences are highlighted, *p < .05.

References

Amodio, D. M., & Devine, P. G. (2006). Stereotyping and evaluation in implicit race bias: Evidence for independent constructs and unique effects on behavior. *Journal of Personality and Social Psychology, 91*, 652–661. https://doi.org/10.1037/0022-3514.91.4.652.

Anderson, A. A., Brossard, D., Scheufele, D. A., Xenos, M. A., & Ladwig, P. (2014). The "nasty effect:" online incivility and risk perceptions of emerging technologies. *Journal of Computer-Mediated Communication, 19*, 373–387. https://doi.org/10.1111/jcc4.12009.

Arendt, F. (2013). Dose-dependent media priming effects of stereotypic newspaper articles on implicit and explicit stereotypes. *Journal of Communication, 63*, 830–851. https://doi.org/10.1111/jcom.12056.

Arendt, F., Marquart, F., & Matthes, J. (2015). Effects of right-wing populist political advertising on implicit and explicit stereotypes. *Journal of Media Psychology: Theories, Methods, and Applications, 27*, 178–189. https://doi.org/10.1027/1864-1105/a000139.

Bagozzi, R. P., & Yi, Y. (1988). On the evaluation of structural equation models. *Journal of the Academy of Marketing Science, 16*, 74–94. https://doi.org/10.1177/009207038801600107.

BAMF. (2016). *Das Bundesamt in Zahlen 2015. Nürnberg, Germany: Bundesamt für Migration und Flüchtlinge*.

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology, 71*, 230–244. https://doi.org/10.1037/0022-3514.71.2.230.

Bargh, J. A., & Pietromonaco, P. (1982). Automatic information processing and social perception: The influence of trait information presented outside of conscious awareness on impression formation. *Journal of Personality and Social Psychology, 43*, 437–449.

Batson, C. D., & Powell, A. A. (2003). Altruism and prosocial behavior. In I. B. Weiner (Ed.), *Handbook of psychology* (pp. 463–484). Hoboken, NJ: John Wiley & Sons.

Blair, I. V., & Banaji, M. R. (1996). Automatic and controlled processes in stereotype priming. *Journal of Personality and Social Psychology, 70*, 1142–1163. https://doi.org/10.1037/0022-3514.70.6.1142.

BMJV. (2017). *Act to Improve Enforcement of the law in social Networks (Network Enforcement act)*. Retrieved from: https://www.bmjv.de.

Bodenhausen, G. V., Macrae, C. N., & Sherman, J. W. (1999). On the dialectics of discrimination. Dual processes in social stereotyping. In S. Chaiken, & Y. Trope (Eds.), *Dual-process Theories in social psychology* (pp. 271–290). New York, NY: Guilford Press.

Braddock, K. (2016). Towards a guide for constructing and disseminating counternarratives to reduce support for terrorism. *Studies in Conflict & Terrorism, 39*, 381–404. https://doi.org/10.1080/1057610X.2015.1116277.

Chen, G. M. (2017). *Online incivility and public debate: Nasty talk*. London, UK: Palgrave Macmillan.

Chen, G. M., & 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. https://doi.org/10.1080/08838151.2016.1273922.

Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication, 64*, 658–679. https://doi.org/10.1111/jcom.12104.

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in Experimental Social Psychology, 40*, 61–149. https://doi.org/10.1016/S0065-2601(07)00002-0.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56*, 5–18. https://doi.org/10.1037/0022-3514.56.1.5.

Dijksterhuis, A., & Bargh, J. A. (2001). The perception-behavior expressway: Automatic effects of social perception on social behavior. *Advances in Experimental Social Psychology, 33*, 1–40. https://doi.org/10.1016/S0065-2601(01)80003-4.

Dixon, T. L. (2008). Crime news and racialized beliefs: Understanding the relationship between local news viewing and perceptions of African Americans and Crime. *Journal of Communication, 58*, 106–125. https://doi.org/10.1111/j.1460-2466.2007.00376.x.

Domingo, D. (2011). Managing audience participation: Practices, workflows and strategies. In J. B. Singer, A. Hermida, D. Domingo, A. Heinonen, S. Paulussen, T. Quandt, et al. (Eds.), *Participatory journalism: Guarding open Gates at online newspapers* (pp. 76–95). Malden, MA: Wiley-Blackwell.

Domke, D. (2001). Racial cues and political ideology. *Communication Research, 28*, 772–801. https://doi.org/10.1177/009365001028006003.

Dovidio, J. F., & Gaertner, S. L. (1993). Stereotypes and evaluative intergroup bias. In D. M. Mackie, & D. L. Hamilton (Eds.), *Affect,Ccognition, and stereotyping* (pp. 167–193). San Diego, CA: Academic Press.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149.

Fiske, S. T. (1989). Examining the role of intent: Toward understanding its role in stereotyping and prejudice. In J. S. Uleman, & J. A. Bargh (Eds.), *Unintended thought* (pp. 253–283). New York: Guilford Press.

Flemming, D., Cress, U., & Kimmerle, J. (2017). Processing the scientific tentativeness of medical research: An experimental study on the effects of research news and user comments in online media. *Science Communication, 39*, 745–770. https://doi.org/10.1177/1075547017738091.

Fornell, C., & Larcker, D. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*, 39–50. https://doi.org/10.1177/002224378101800104.

Gaertner, S. L., Dovidio, J. F., Rust, M. C., Nier, J. A., Banker, B. S., Ward, C. M., … Houlette, M. (1999). Reducing intergroup bias: Elements of intergroup cooperation. *Journal of Personality and Social Psychology, 76*, 388–402. https://doi.org/10.1037//0022-3514.76.3.388.

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). *Countering online hate speech.* Paris, FR: UNESCO.

Garcia, S. M., Weaver, K., Moskowitz, G. B., & Darley, J. M. (2002). Crowded minds: The implicit bystander effect. *Journal of Personality and Social Psychology, 83*, 843–853. https://doi.org/10.1037/0022-3514.83.4.843.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132*, 692–731. https://doi.org/10.1037/0033-2909.132.5.745.

Gervais, B. T. (2014). Incivility online: Affective and behavioral reactions to uncivil political posts in a web-based experiment. *Journal of Information Technology & Politics, 12*, 167–185. https://doi.org/10.1080/19331681.2014.997416.

Gilliam, F. D., & Iyengar, S. (2000). Prime suspects: The influence of local television news on the viewing public. *American Journal of Political Science, 44*, 560–573. https://doi.org/10.2307/2669264.

Greenblatt, J. (2015). When hateful speech leads to hate crimes. Huffington Post, 21st August 2015 https://www.huffingtonpost.com/jonathan-greenblatt/when-hateful-speech-leads_b_8022966.html.

Greenwald, A. G., Banaji, M., Rudman, L., Farnham, S., Nosek, B., & Mellot, D. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review, 109*, 3–25. https://doi.org/10.1037/0033-295X.109.1.3.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*, 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology, 85*, 197–216. https://doi.org/10.1037/0022-3514.85.2.197.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the implicit association test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97*, 17–41. https://doi.org/10.1037/a0015575.

Harteveld, E., Schaper, J., De Lange, S. L., & Van Der Brug, W. (2018). Blaming brussels? The impact of (news about) the refugee crisis on attitudes towards the EU and national politics. *Journal of Communication and Media Studies: Journal of Common Market Studies, 56*, 157–177. https://doi.org/10.1111/jcms.12664.

Higgins, E. T. (1996). Knowledge activation: Accessibility, applicability, and salience. In E. T. Higgins, & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 133–168). New York, NY: Guilford Press.

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the Implicit Association Test and explicit self-report measures. *Personality and Social Psychology Bulletin, 31*, 1369–1385. https://doi.org/10.1177/0146167205275613.

Hsueh, M., Yogeeswaran, K., & Malinen, S. (2015). "Leave your comment below": Can biased online comments influence our own prejudicial attitudes and behaviors? *Human Communication Research, 41*, 557–576. https://doi.org/10.1111/hcre.12059.

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

Hurwitz, J., & Peffley, M. (2005). Playing the race card in the post-Willie Horton era: The impact of racialized code words on support for punitive crime policy. *Public Opinion Quarterly, 69*, 99–112. https://doi.org/10.1093/poq/nfi004.

Köcher, R. (2016). *Vertrauenskrise der Medien? [A crisis of media trust]*. Berlin: Presentation at the VDZ Publishers' Summit, 8th November 2016.

Lee, E. J., & Jang, Y. J. (2010). What do others' reactions to news on internet portal sites tell us? Effects of presentation format and readers' need for cognition on reality perception. *Communication Research, 37*, 825–846. https://doi.org/10.1177/0093650210376189.

LfM. (2017). Hate speech. Retrieved from https://www.medienanstalt-nrw.de/.

Lobo, S. (2015). *Wie aus Netzhass Gewalt wird und was dagegen hilft [How online hate becomes violence und what to do about it]*. Spiegel Online, 19th August 2015 http://www.spiegel.de/netzwelt/netzpolitik/netzhass-und-gewalt-was-man-dagegen-tun-kann-lobo-kolumne-a-1048799.html.

Macnair, L., & Frank, R. (2017). Voices against extremism: A case study of a community-based CVE counter-narrative campaign. *Journal for Deradicalization, 10*, 147–174.

Mutz, D. C., & Reeves, B. (2005). The new videomalaise: Effects of televised incivility on political trust. *American Political Science Review, 99*, 1–15.

Newman, N., Fletcher, R., Levy, D. A. L., & Nielsen, R. K. (2016). Reuters institute digital news report *2016*. Retrieved from http://www.digitalnewsreport.org/.

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., et al. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology, 18*, 36–88. https://doi.org/10.1080/10463280701489053.

Olson, M. A. (2009). Measures of prejudice. In T. D. Nelson (Ed.), *Handbook of prejudice, stereotyping and discrimination* (pp. 367–386). New York: Psychology Press.

Olson, M. A., & Fazio, R. H. (2006). Reducing automatically activated racial prejudice through implicit evaluative conditioning. *Personality and Social Psychology Bulletin, 32*(4), 421–433. https://doi.org/10.1177/0146167205284004.

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

Power, J. G., Murphy, S. T., & Coover, G. (1996). Priming prejudice: How stereotypes and counter-stereotypes influence attribution of responsibility and credibility among ingroups and outgroups. *Human Communication Research, 23*, 36–58. https://doi.org/10.1111/j.1468-2958.1996.tb00386.x.

Roskos-Ewoldsen, D. R., Klinger, M. R., & Roskos-Ewoldsen, B. (2007). Media priming: A meta-analysis. In R. W. Preiss (Ed.), *Mass media effects research. Advances through meta-analysis* (pp. 53–80). New York: Lawrence Erlbaum.

Roskos-Ewoldsen, D. R., Roskos-Ewoldsen, B., & Dillmann Carpentier, F. R. (2002). Media priming: A synthesis. In J. Bryant, & M. B. Oliver (Eds.), *Media effects: Advances in theory and research* (pp. 97–120). New York, NY: Routledge.

Rösner, L., Winter, S., & Krämer, N. C. (2016). Dangerous minds?: Effects of uncivil online comments on aggressive cognitions, emotions, and behavior. *Computers in Human Behavior, 58*, 461–470. https://doi.org/10.1016/j.chb.2016.01.022.

Rowe, I. (2015). Civility 2.0: A comparative analysis of incivility in online political discussion. *Information, Communication & Society, 18*, 121–138. https://doi.org/10.1080/1369118X.2014.940365.

Schemer, C. (2012). The influence of the news media on stereotypic attitudes toward immigrants in a political campaign. *Journal of Communication, 62*, 739–757. https://doi.org/10.1111/j.1460-2466.2012.01672.x.

Srull, T. K., & Wyer, R. S. (1980). Category accessibility and social perception: Some implications for the study of person memory and interpersonal judgments. *Journal of Personality and Social Psychology, 38*, 841–856. https://doi.org/10.1037/0022-3514.38.6.841.

Stromer-Galley, J. (2007). Measuring deliberation's content: A coding scheme. *Journal of Public Deliberation, 3*(1). Article 12. Available at: https://www.publicdeliberation.net/jpd/vol3/iss1/art12.

Stroud, N. J., Scacco, J. M., Muddiman, A., & Curry, A. L. (2015). Changing deliberative norms on news organizations' Facebook sites. *Journal of Computer-Mediated Communication, 20*, 188–203. https://doi.org/10.1111/jcc4.12104.

Stroud, N. J., Van Duyn, E., & Peacock, C. (2016). *News commenters and news comment readers*. Engaging News Project. Retrieved from https://engagingnewsproject.org.

Su, L. Y. F., Xenos, M. A., Rose, K. M., Wirz, C., Scheufele, D. A., & Brossard, D. (2018). Uncivil and personal? Comparing patterns of incivility in comments on the Facebook pages of news outlets. *New Media & Society, 20*(10), 3678–3699. https://doi.org/10.1177/1461444818757205.

Tajfel, H., & Turner, J. C. (1986). The social identity theory of intergroup behavior. In S. Worchel, & W. G. Austin (Eds.), *Psychology of intergroup Relations* (pp. 7–24). Chicago, IL: Nelson-Hall.

Valkenburg, P. M., & Peter, J. (2013). The differential susceptibility to media effects model. *Journal of Communication, 63*, 221–243. https://doi.org/10.1111/jcom.12024.

Van Baaren, R. B., Holland, R. W., Kawakami, K., & Van Knippenberg, A. (2004). Mimicry and prosocial behavior. *Psychological Science, 15*, 71–74. https://doi.org/10.1111/j.0963-7214.2004.01501012.x.

Winter, S., Brückner, C., & Krämer, N. C. (2015). They came, they liked, they commented: Social influence on Facebook news channels. *Cyberpsychology, Behavior, and Social Networking, 18*, 431–436. https://doi.org/10.1089/cyber.2015.0005.

Ziegele, M., & Jost, P. B. (2016). Not funny?: The effects of factual versus sarcastic journalistic responses to uncivil user comments. *Communication Research*, 1–30. https://doi.org/10.1177/0093650216671854.

Ziegele, M., Koehler, C., & 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*, (4), 636–653. https://doi.org/10.1080/08838151.2018.1532430.

Ziegele, M., Naab, T., & Jost, P. (2019). Lonely together? Identifying the determinants of collective corrective action against uncivil comments. *New Media & Society*, 1–21. https://doi.org/10.1177/1461444819870130.

Ziegele, M., & Quiring, O. (2017). The discussion value of online news: How news story characteristics affect the deliberative quality of user discussions in SNS comment sections. In *Paper presented at the 67th Annual Conference of the International communication association (ICA), may 25-29, san Diego, USA*.