MUNI FSS

JOURNAL OF PSYCHOSOCIAL RESEARCH ON CYBERSPACE
CYBERPSYCHOLOGY

Jia, Y., & Schumann, S. (2025). Tackling hate speech online: The effect of counter-speech on subsequent bystander behavioral intentions. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace, 19*(1), Article 4. https://doi.org/10.5817/CP2025-1-4

Tackling Hate Speech Online: The Effect of Counter-Speech on Subsequent Bystander Behavioral Intentions

Yue Jia[1] *& Sandy Schumann*[2]

[1] Social Research Institute, University College London, London, UK
[2] Department of Security and Crime Science, University College London, London, UK

Abstract

Counter-speech is considered a promising tool to address hate speech online, notably, by promoting bystander reactions that could attenuate the prevalence or further dissemination of hate. However, it remains unclear which types of counter-speech are most effective in attaining these goals and which might backfire. Advancing the literature, we examined the effect of four types of counter-speech (i.e., educating the perpetrator, calling on others to intervene, diverting the conversation, and abusing the perpetrator) on a range of bystander behavioral intentions in an experimental study (N = 250, UK-based adults). Overall, counter-speech did not affect bystanders' subsequent responses to hate speech. Having said this, as expected, diversionary counter-speech increased intentions to ignore hate speech, which suggests unintended consequences. The study illustrates that counter-speech may not be sufficiently impactful in regulating bystanders' reactions to hate speech online.

Keywords: hate speech; bystander intervention; bystander reaction; counter-speech; online experiment

Editorial Record

First submission received:
April 6, 2024

Revisions received:
October 8, 2024
December 19, 2024

Accepted for publication:
December 19, 2024

Editor in charge:
Lenka Dedkova

Introduction

Survey studies indicate that a substantial number of internet users have been the target of or have observed hate speech online (Bergmann & Baier, 2018; Oksanen et al., 2014; Pacheco & Melhuish, 2018; Räsänen et al., 2016). To address this concern, technology companies have implemented algorithm-driven protocols to detect, remove, or quarantine hate speech. Complementing those measures, direct user responses—counter-speech—are considered a valuable tool for regulating hate speech (Bartlett & Krasodomski-Jones, 2015; Howard, 2021; W. Zhu & Bhat, 2021). Notably, in addition to changing perpetrators' behavior (Garland et al., 2020; Hangartner et al., 2021; Mathew et al., 2018), counter-speech can facilitate subsequent bystander responses that have the potential to attenuate the prevalence, spread, and adverse impact of hate speech (Garland et al., 2022; Obermaier et al., 2021). However, little is known about the boundary conditions of this effect (see Lasser et al., 2023).

Specifically, thus far, it has not been acknowledged systematically that not all counter-speech is effective or, relatedly, that certain types of counter-speech could have unintended consequences by promoting behavior that facilitates the dissemination of hate speech (e.g., sharing hate speech). The present study advances this literature. We compared the influence of four common types of counter-speech (i.e., educating the perpetrator that their hateful views are wrong, directly calling on others to intervene, diverting the conversation, and abusing the perpetrator) on a range of immediate bystander behavioral intentions that could reduce or enhance the proliferation and negative impact of hate speech. To examine our hypotheses, we conducted an experimental

study, complementing insights gained from previous analyses of social media data (Friess et al., 2021; Garland et al., 2022; Lasser et al., 2023) that only captures directly observable behavior.

Online Hate Speech

For the purpose of this research, we define hate speech as public speech that expresses hate or encourages violence towards a person or group based on protected characteristics such as ethnicity, religion, gender, or sexual orientation (Cambridge Dictionary, n.d.; Council of Europe, n.d.; Stop Hate UK, n.d.). Online hate speech then refers to hate speech that is posted on or disseminated via social media platforms, in online gaming contexts, or through other means of computer-mediated communication (Corazza et al., 2020; Ortiz, 2019; Rieger et al., 2021). Different modalities, such as memes, text, and videos, can be employed to convey online hate speech.

Approximately .001% to 1% of content on mainstream social media platforms is estimated to be classified as hate speech; on fringe platforms such as 4chan and Gab, the proportion is likely between 5% to 8% (Vidgen et al., 2019). In the 2017–2018 Hate Crimes (England and Wales) report, 1,605 hate crimes were considered to contain online elements, accounting for 2% of the total incidents (Home Office, 2018). These relatively low figures may not appear as cause for concern. However, multiple surveys have found that 50% to 80% of teenagers have been exposed to online hate material, and around 20% of teenagers have been the direct victims (Oksanen et al., 2014; Winiewski et al., 2017).

The internet may promote the expression and spread of hate speech for several reasons. First, users can remain anonymous (Mondal et al., 2018), which might evoke a sense of disinhibition such that perpetrators feel less restrained and express themselves more violently (Suler, 2004). Second, the instantaneous nature of the internet encourages impulsive hate speech (Brown, 2018), representing the most common type of hate crime offender (McDevitt et al., 2002). Third, one-click features such as “share” and “retweet” make it easier to disseminate hate speech to a wide audience, affording repeated victimization (see Benigni et al., 2017; Veilleux-Lepage, 2016). Moreover, social bots that automatically retweet posts without verifying facts could further accelerate the dissemination of hate speech (Ferrara et al., 2016). In fact, an analysis of Twitter conversations during the COVID-19 pandemic showed that a higher prevalence of bots was associated with more incidents of online hate (Uyheng & Carley, 2020). Lastly, negative information, such as hate speech, spreads more easily and quickly online (Maarouf et al., 2022). Tsugawa and Ohsaki (2015) illustrated this point and found that the reposting volume of negative news was 1.2–1.6 times that of positive and neutral news, and negative news spread 1.25 times faster.

Countermeasures and Bystander Intervention

Measures to counter online hate speech involve several stakeholders. A large number of social media platforms have implemented technical solutions to detect problematic content, either by enabling internet users’ reporting/flagging (Crawford & Gillespie, 2016) or by introducing algorithmic methods such as hashing and classification (Farid, 2021; Vidgen et al., 2021). Once hate speech is identified, it is deleted. Alternatively, quarantine and deplatforming reduce the supply of online hate speech by blocking detected content temporarily (Ullmann & Tomalin, 2020) or by removing the accounts that have disseminated the material (Rogers, 2020; see also Copland, 2020; Jhaver et al., 2021). Social media platforms’ efforts to detect and remove or quarantine hate speech online are, to some extent, the result of regulatory pressures (Chetty & Alathur, 2018). For example, the UK’s Public Order Act 1986 stipulates that people who threaten, abuse, and insult others can be sentenced to up to six months in prison (UK Legislation, n.d.). Moreover, new school speech regulations, such as the newly passed 2023 UK Online Safety Act or the German Network Enforcement Act (NetzDG law), require technology companies to remove hate speech within certain time frames (Balkin, 2017; Bundesministerium der Justiz, 2017; UK Legislation, 2023).

Technical and legal countermeasures, though advantageous in several ways, have inherent limitations. They are often criticized for restricting free speech (Human Rights Watch, 2018), and their effectiveness remains unproven, among others, due to a lack of systematic compliance (Copland, 2020; Griffin, 2022). Complementing those activities, bottom-up approaches rely on the intervention of internet users who encounter hate speech online, that is, bystanders. One form of bystander intervention is counter-speech, that is, a direct reply to hate speech online; it includes but is not limited to, presenting facts that contest a hateful comment, denouncing hate speech, warning of the consequences of hate speech, distraction, attacking/insulting the perpetrators, and showing empathy for the victims (Benesch et al., 2016; Cepollaro et al., 2023; Mathew et al., 2019; Obermaier et al., 2021).

Previous research has identified situational and personal factors that enhance the willingness to actively intervene in incidents of online incivility, including hate speech. A lower number of other bystanders (Obermaier et al., 2016), higher severity of the incident (Bastiaensens et al., 2014), and a lower risk of harm (Thomas et al., 2012) were associated with a higher willingness to take actions, such as expressing counter-speech. Higher levels of empathy (Machackova et al., 2015), self-efficacy (DeSmet et al., 2016), and self-control (Erreygers et al., 2016), as well as lower levels of prior victimization experiences (Barlińska et al., 2013) and moral disengagement (DeSmet et al., 2014) also facilitated people's tendencies to intervene.

Counter-speech is a promising tool to reduce the spread of hate speech online as it can affect the subsequent behavior of perpetrators (Garland et al., 2020; Hangartner et al., 2021; Mathew et al., 2018; Miškolci et al., 2020) and other bystanders. Speaking especially to the latter point, analyses of political conversations on Twitter in Germany (Garland et al., 2022) and counter-speech against online Islamophobic hate speech (Obermaier et al., 2021) found that counter-speech tended to facilitate more subsequent counter-speech. However, not all types of counter-speech attain the same effect. Friess et al. (2021) concluded, based on Facebook comments extracted from the German action group #ichbinhier, that counter-speech written by in-group members inspired more deliberative comments than counter-speech posted by out-group members. Civil counter-speech encouraged participants to engage in on-topic and civil discussions, and uncivil comments promoted more meta-communication but did not predict a significant growth in uncivil responses (Han et al., 2018). Additionally, Lasser and colleagues (2023) showed that offering simple opinions, without insults, and using sarcasm predicted a lower prevalence of hate, toxicity, and extremism in the long term; having said this, a short-term increase in the problematic discourse was observed. Importantly, counter-speech that raised the salience of group divisions was related to a rise in hate, toxicity, and extremism (Lasser et al., 2023). In other words, some counter-speech can have unintended adverse consequences. To date, however, it has not yet been documented systematically which types of counter-speech elicit desirable bystander behavior that promises to prevent the proliferation or reduce the negative impact of hate speech and which counter-speech enhances counter-productive responses.

The Present Study

The present study addresses this gap in the literature. In doing so, we focused on the implications of exemplars of four common types of counter-speech. We selected the latter based on the results of a content analysis of around 2,000 hate conversations, 6,000 instances of counter-speech, and 1,000 subsequent bystander reactions on X (former Twitter) that identified the following types of counter-speech: attacking the perpetrator, fact-based educational speech, emotion-based educational speech, simple disagreement without trying to persuade, and off-topic/neutral speech (Jia & Schumann, 2024). The four types of counter-speech assessed in this study represent manifestations of the four most common types of counter-speech on X, namely: a) educating the perpetrator that their hateful views are wrong; b) directly calling on others to intervene in hate speech; c) diverting the conversation; and d) abusing the perpetrator. As outcomes of interest, we considered bystanders' intentions to engage in eight behaviors immediately after observing one instance of counter-speech. Specifically, we assessed intentions to report hate speech, educate perpetrators that their comments are unacceptable, and comfort victims, which are generally seen as desirable bystander reactions (DeSmet et al., 2016; Macaulay et al., 2022) that can contribute to building supportive and constructive online discourse (Crawford & Gillespie, 2016; DeSmet et al., 2016; Lasser et al., 2023). We further investigated the influence of counter-speech on intentions to post a comment expressing a similar position as the perpetrator or sharing hate speech, that is, undesirable reactions that further accelerate the spread of online hate and escalate incivility (Van Cleemput et al., 2014). Additionally, we assessed intentions to post an offensive comment condemning the perpetrators; this response might aim to fight hatred but is likely to evoke further toxic and uncivil speech (Lasser et al., 2023). Furthermore, we examined the likely most common bystander behavior: inaction (Van Cleemput et al., 2014). Ignoring hate speech (i.e., inaction) does not contribute to escalating hate but equally does not help combat incivility or support victims (Macaulay et al., 2022; Van Cleemput et al., 2014). Moreover, inaction could be understood as "silent approval" of hate speech (DeSmet et al., 2016; Jeyagobi et al., 2022, p. 2; Song & Oh, 2018), although its precise implications are not yet well understood (e.g., Friess et al., 2021; Garland et al., 2022; Lasser et al., 2023). Finally, we investigated the effect of counter-speech on intentions to post comments that are unrelated to the original topic, trying to divert the conversation; those could be perceived as online trolling and elicit uncivil responses (Cheng et al., 2017) or an attempt to reduce the negative impact of hate speech (see Barberá et al., 2022).

Previous research suggests distinct effects of certain types of counter-speech on particular bystander reactions. More precisely, educational counter-speech demonstrates caring and a recognition of the injustice experienced by victims (Hoffman, 2014). Several studies have recognized that such expressions of empathy promote constructive bystander behavior, for instance, helping victims (Freis & Gurung, 2013; Van Cleemput et al., 2014). Relatedly, educational counter-speech has also been found to help shape norms of solidarity that could increase the likelihood of reporting counter-speech (Kunst et al., 2021). In addition, if educational counter-speech is civil, it should promote further civil bystander responses (Molina & Jennings, 2018). Taken together, we propose that posting counter-speech designed to educate perpetrators that their hateful views are wrong, without using insults, strengthens bystanders' intentions to a) report hate speech, b) post a comment to educate the perpetrator, and c) comfort the victims (**H1**).

By contrast, counter-speech that directly calls on other users to intervene in hate speech might promote inaction. Calls for action are often expressed in an assertive tone, a strategy that is also used in advertising (i.e., assertive ads), such as "You must try our …" and "Only you can …" (Kim et al., 2017, p. 551). It has been demonstrated that assertive ads can be effective in attracting attention; however, they are perceived as manipulative, weakening their persuasiveness and increasing consumer reactance (Edwards et al., 2002; Quick & Stephenson, 2007; Zemack-Rugar et al., 2017; see also Miron & Brehm, 2006; Steindl et al., 2015). Accordingly, we postulate that posting counter-speech that directly calls on others to intervene in hate speech increases bystanders' intentions to ignore hate speech and remain inactive (**H2**).

Diversionary counter-speech, that is, talking about a topic that is not related to hate speech incidents, has not been explored in previous research. However, campaigns that aim to encourage the public to take action when observing hate speech or crime incidents often recommend deflecting the conversation (Media Smarts, n.d.). Distraction is also used for political propaganda to reduce attention to and stifle debate about controversial or unwanted topics (Barberá et al., 2022; King et al., 2017). Based on this evidence, we speculate that diversionary counter-speech will reduce bystanders' attention to hate speech, translating into stronger intentions to a) ignore hate speech and b) post an unrelated comment (**H3**).

Lastly, we postulate that counter-speech that abuses the perpetrator has negative implications and promotes undesirable bystander responses (Lasser et al., 2023). The literature on the contagion effect (Buerger, 2022; Kramer et al., 2014) and the broken windows theory (Wilson & Kelling, 1982) highlights that such uncivil behavior is contagious and may lead others to view offensive language, including such that targets the perpetrator, as acceptable. In addition, uncivil speech was shown to increase readers' hostile cognition and the posting of further hate speech (Benesch et al., 2016; Rösner et al., 2016; however, see Han & Brazeal, 2015 for opposing findings). Therefore, we believe that posting counter-speech that abuses the perpetrator increases bystanders' intentions to subsequently a) post a comment to express a similar position as the perpetrator, b) post a comment condemning the perpetrator with offensive words, and c) share hate speech (**H4**).

Methods

The study was approved by the authors' departmental ethics committee. All participants read the Participant Information Form and agreed to a Consent Form, which included an agreement to share data for further scientific research. The study material and data are available in the associated OSF repository here.

Design and Participants

To test the hypotheses, we conducted a one-factorial between-subject experiment online. All participants were randomly assigned to one of five conditions (i.e., four experimental groups, G1–G4, representing different types of counter-speech, and one control group, G5). After removing four participants who failed the attention check, N = 246 valid responses remained for analysis. The sample consisted of n = 148 women (60.16%) and n = 98 men (39.84%), and the mean age was M = 35.59 (SD = 12.80, Min = 18, Max = 74). Regarding their ethnicity, there were n = 119 Asian-Chinese; n = 5 Mixed/Chinese, n = 78 White, n = 37 Asian-Other, n = 5 Black, n = 2 Mixed or Other ethnicity participants included in the sample. Among them, n = 213 (86.59%) claimed to have previously been bystanders of hate speech online, n = 45 (18.29%) reported having been direct victims, and n = 8 (3.25%) admitted to having posted hate speech online.

Procedure

Data were collected in July 2022 using the opt-in access panel Prolific Academic. The average survey completion time was six minutes. After providing informed consent, participants were asked to report their gender, age, and ethnicity. Next, we provided definitions of "hate speech" and "counter-speech" and captured information on participants' online activities and experiences. Then, based on the experimental condition to which they were assigned, participants read a tweet thread containing an initial post, a racist hate speech reply, and a counter-speech response. In the control condition, no counter-speech was shown in response to the hate speech comment, as even a post on an unrelated topic would constitute counter-speech (see condition "diversionary counter-speech"). Following a manipulation check, participants reported their willingness to engage in eight different bystander reactions.

Materials

Experimental Stimuli

The experimental stimuli included screenshots of tweets that we had designed[1]. At the time of data collection, anti-Chinese racist hate speech triggered by the COVID-19 pandemic was highly prevalent online (H. Zhu, 2020); we, therefore, developed stimulus material that included anti-Chinese hate speech. All source material for the tweets was derived from open-source hate speech datasets (De Gibert et al., 2018; Kennedy et al., 2020; Röttger et al., 2021; Vidgen et al., 2020). The authors' names were replaced by the letters A, B, and C appearing in different colors. Participants in each experimental group (i.e., G1–G4) read tweets including A's initial post, B's racist speech, and C's counter-speech (1). In the control group (i.e., G5), the stimulus included only A's initial post and B's racist speech as a response (Figure 2).

Figure 1. *Four Types of Counter-Speech.*

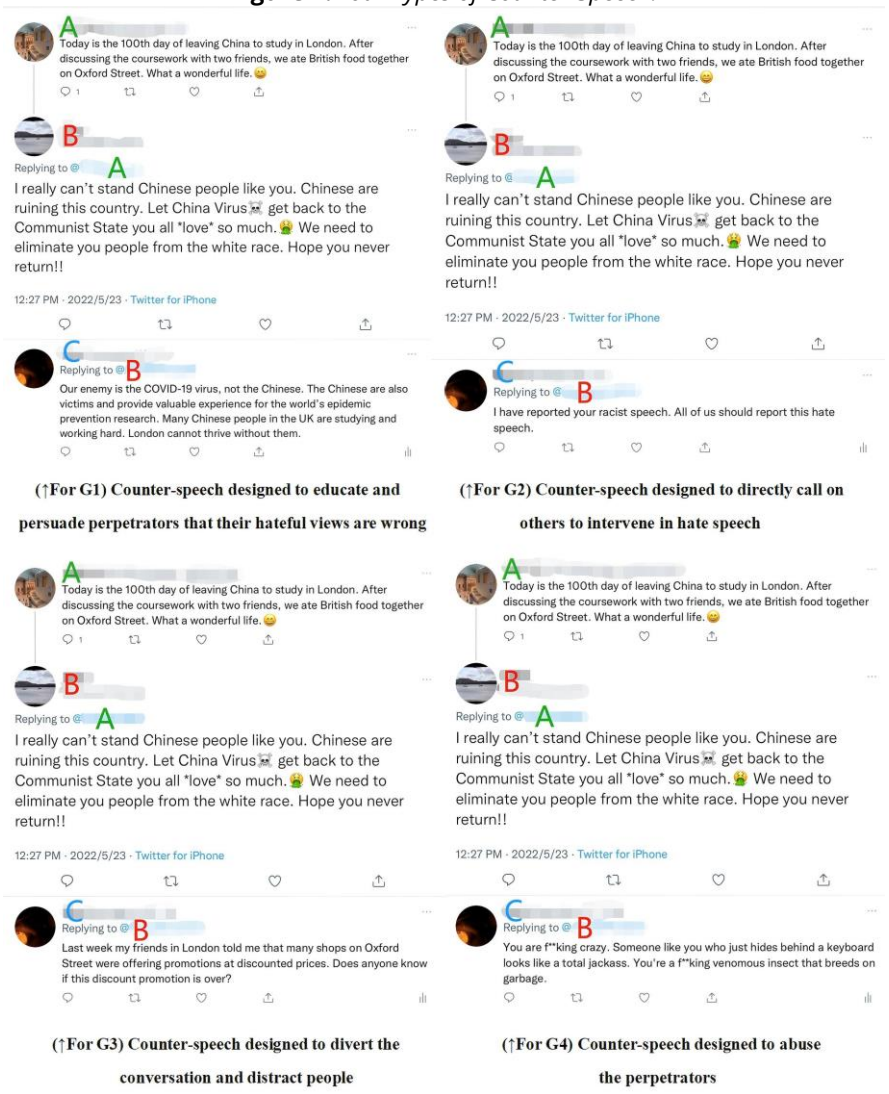

Figure 2. *Tweets Read by G5 (Control Group).*

Measures

To capture the dependent variables, that is, immediate bystander behavioral intentions, participants indicated how likely they were to take the following eight actions (1 = *strongly unlikely*, 5 = *strongly likely*): *Report B's racist tweet to the platform*; *Post a comment to educate and persuade B*; *Post a comment to comfort and show empathy for A*; *Post an unrelated comment*; *Post a comment to express a similar position as B*; *Share B's tweets by retweeting*; *Ignore these tweets and continue browsing*; *Post a comment condemning B with potentially offensive words*.

Social desirability bias negatively affects the authenticity of answers (Krumpal, 2013). We expected that reminding participants of the confidentiality and anonymity of the questionnaire would reduce this bias to a certain extent. Therefore, participants were asked to indicate their behavior intentions once more after reading a note that stated: *Any information you provide will be kept confidential. You may maintain or change your choices*. All analyses described below were conducted using the data of this second measurement of the dependent variable. To assess the impact of the reminder, we conducted Wilcoxon signed-rank tests for each paired response (i.e., eight pairs in total). Bonferroni correction to alpha levels was applied. The findings suggested that after a reminder of the confidentiality and anonymity of the questionnaire, participants' responses were significantly higher for the dependent variable *Post a comment to educate and persuade B* ($Z = -3.187$, $p = .001$, $r = .144$) and significantly lower for the dependent variable *Post a comment to express a similar position as B* ($Z = -2.865$, $p = .004$, $r = .129$). No significant differences were found for the remaining six dependent variables (all $p > .006$). These results suggest that the reminder made people reevaluate their assessment to some extent and was, thus, successful.

Participants' internet use patterns were assessed by asking how often they used social media to "browse information", "comment on posts", and "share information" (1 = *never*, 5 = *very often*). Additionally, participants indicated whether they had ever been a bystander, victim, or perpetrator of hate speech online (*Yes*/*No*). Lastly, as a manipulation check, we asked participants in the four experimental groups the following question: *How strongly do you agree or disagree with the following statements about C's counter-speech?*. Five answers were provided: *C aims to abuse B*; *to educate and persuade B*; *to divert the conversation and distract people*; *to directly call on others to intervene in B's hate speech*; *to express humor* (1 = *strongly disagree*, 5 = *strongly agree*).

Results

Data analysis was performed using IBM SPSS 29.

Sensitivity Analyses

Sensitivity analyses were performed using G*Power 3.1.9.7 (Perugini et al., 2018). As detailed below, the hypotheses tests included independent samples *t*-tests[2]. Setting adjusted $\alpha = .017/0.050/0.025/0.017$ for *Hypothesis* 1/2/3/4 and power $(1-\beta) = .80$, the analyses identified the effect size of Cohen's $d = .655/.569/.625/.666$ for each hypothesis. Thus, the hypotheses tests could detect medium to large effects.

The exploratory analyses pertained to Kruskal-Wallis tests. The results showed that for conducting a Kruskal-Wallis test[3], given the sample of $N = 246$ with five groups, setting adjusted $\alpha = .006$ and power $(1-\beta) = .80$, the minimum effect size that could be reliably detected was $f = .276$—a medium effect.

Descriptive Results

The average frequencies of participants using social media to browse information, comment on posts, and share information were $M = 4.16$ ($SD = .96$), 2.54 ($SD = .98$), and 2.60 ($SD = 1.02$), respectively. That is, participants can be described as primarily passive, and only moderately active, social media users (Trifiro & Gerson, 2019). Table 1 shows the means of all bystander behavioral intentions across conditions. Overall, participants were most likely to report or ignore hate speech, and least likely to share hate speech by retweeting or to express a similar position as the perpetrator.

Table 1. *Means of Eight Bystander Behavioral Intentions Across the Five Conditions.*

Experimental condition	Report B's racist tweet to the platform	Post a comment to educate and persuade B	Post a comment to comfort and show empathy for A	Post an unrelated comment	Post a comment to express a similar position as B	Share B's tweets by retweeting	Ignore these tweets and continue browsing	Post a comment condemning B with potentially offensive words
	M (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
G1 (...educate...) (*n* = 50)	3.22 (1.42)	2.14 (1.16)	2.76 (1.35)	1.48 (0.79)	1.22 (0.55)	1.32 (0.65)	4.10 (0.95)	1.48 (0.95)
G2 (...call on...) (*n* = 49)	3.71 (1.37)	2.10 (1.14)	2.76 (1.42)	1.35 (0.63)	1.10 (0.31)	1.14 (0.50)	3.45 (1.27)	1.84 (1.07)
G3 (...divert...) (*n* = 50)	3.60 (1.39)	1.84 (1.06)	2.76 (1.27)	1.70 (0.95)	1.04 (0.20)	1.22 (0.65)	4.16 (0.82)	1.48 (0.81)
G4 (...abuse...) (*n* = 47)	3.30 (1.41)	1.75 (0.97)	2.34 (1.26)	1.70 (0.98)	1.17 (0.52)	1.23 (0.60)	3.94 (1.17)	1.70 (1.10)
G5 (control) (*n* = 50)	3.70 (1.46)	2.04 (1.11)	2.52 (1.20)	1.52 (0.76)	1.22 (0.58)	1.46 (0.99)	3.62 (1.10)	2.06 (1.13)

Manipulation Check

One-sample *t*-tests with the test value of $M = 4$ ($H0$: $M \geq 4$, thus representing the average answer options *Agree* or *Strongly agree*) were performed for the manipulation check, separately in each of the four experimental conditions. As shown in Table 2, all target answers (i.e., statements that described what a tweet ought to convey) could not reject the null hypothesis ($M \geq 4$), and all other answers rejected the null hypothesis. In other words, participants perceived the counter-speech stimuli in line with their intended purpose.

Table 2. *Results of the Manipulation Check.*

Experimental condition	Answer (#: target answer; C: bystander; B: perpetrator)	*M*	*SD*	*p* (*H0*: $M \geq 4$)
G1 (…educate…) (*n* = 50)	C aims to abuse B	1.52	0.95	< .001*
	C aims to educate and persuade B #	4.54	0.58	> .999
	C aims to divert the conversation and distract people	2.34	1.19	< .001*
	C aims to directly call on others to intervene in B's hate speech	2.22	1.22	< .001*
	C aims to express humor	1.38	0.75	< .001*
G2 (…call on…) (*n* = 49)	C aims to abuse B	1.45	1.04	< .001*
	C aims to educate and persuade B	2.61	1.32	< .001*
	C aims to divert the conversation and distract people	1.80	1.08	< .001*
	C aims to directly call on others to intervene in B's hate speech #	4.57	0.79	>.999
	C aims to express humor	1.12	0.44	< .001*
G3 (…divert…) (*n* = 50)	C aims to abuse B	1.18	0.48	< .001*
	C aims to educate and persuade B	1.42	0.84	< .001*
	C aims to divert the conversation and distract people #	4.02	0.98	.557
	C aims to directly call on others to intervene in B's hate speech	1.56	0.99	< .001*
	C aims to express humor	2.14	1.11	< .001*
G4 (…abuse…) (*n* = 47)	C aims to abuse B #	3.68	1.29	.048
	C aims to educate and persuade B	2.13	1.15	< .001*
	C aims to divert the conversation and distract people	2.32	1.09	< .001*
	C aims to directly call on others to intervene in B's hate speech	2.60	1.35	< .001*
	C aims to express humor	1.68	1.02	< .001*

Note. *$p < .010$ after Bonferroni correction ($\alpha = .050/5$); G1–G4 stand for Group 1 to Group 4.

Hypotheses Tests

We stipulated distinct effects of the four types of counter-speech on particular outcomes, that is, bystander behavioral intentions. In examining these effects, we conducted several independent samples Welch *t*-tests. Each test compared an experimental group with the control group. Table 3 shows the results of all tests.

Hypothesis 1 proposed that posting counter-speech to educate perpetrators, without using insults, increases intentions to subsequently a) report hate speech, b) post a comment to educate and persuade the perpetrator, and c) post a comment to comfort and show empathy for the victims. Findings showed that Hypothesis 1 was rejected. We further postulated that posting counter-speech designed to directly call on others to intervene in hate speech enhances intentions to subsequently ignore hate speech (Hypothesis 2). Hypothesis 2 was also not supported. Hypothesis 3 argued that posting counter-speech designed to divert the conversation increases bystanders' intentions to a) ignore hate speech, and b) post an unrelated comment. The results showed that Hypothesis 3b was rejected but that Hypothesis 3a was supported. Lastly, Hypothesis 4 postulated that posting counter-speech that abuses the perpetrator strengthens intentions to subsequently a) post a comment to express a similar position as the perpetrator, b) post a comment condemning the perpetrator with offensive words, and c) share hate speech. However, none of these effects were statistically significant; Hypothesis 4 was rejected.

Table 3. *Results of Independent Samples Welch t-Tests.*

Experimental group	Bystander behavioral intention (B: perpetrator; A: victim)	Mean difference (experimental–control)	*t*	*df*	*p*	*d*
G1 (…educate…)	Report B's racist tweet to the platform	−.480	−1.667	98	.099	.333
	Post a comment to educate and persuade B	.100	0.441	98	.660	.088
	Post a comment to comfort and show empathy for A	.240	0.940	97	.349	.188
G2 (…call on…)	Ignore these tweets and continue browsing	−.171	−0.712	94	.478	.143
G3 (…divert…)	Ignore these tweets and continue browsing [a]	.540	2.779	90	.007*	.556
	Post an unrelated comment	.180	1.043	93	.300	.209
G4 (…abuse…)	Post a comment to express a similar position as B	−.050	−0.443	95	.659	.090
	Post a comment condemning B with potentially offensive words	−.358	−1.578	95	.118	.320
	Share B's tweets by retweeting [a]	−.226	−1.366	81	.176	.276

Note. $^{*}p < .025$ after Bonferroni correction; G1–G4: Group 1 to Group 4; [a] equal variances were not assumed due to failed Levene's test for equality of variances.

Exploratory Analyses

Further exploratory analyses included eight Kruskal-Wallis tests in which we compared all five experimental conditions/the control group with respect to each type of bystander behavioral intention as an outcome. After correcting the alpha level for the eight comparisons (α = .006), results suggested no statistically significant between-group differences (Table 4). Having said this, given that Bonferroni corrections are conservative, we found here once again tentative evidence that different types of counter-speech vary in their effects on ignoring hate speech.

Table 4. *Results of Kruskal-Wallis Tests.*

Factor	Bystander behavioral intention (B: perpetrator; A: victim)	*H*	*df*	*p*	η^2
Experimental/control condition	Report B's racist tweet to the Twitter platform	6.419	4	.170	.010
	Post a comment to educate and persuade B	5.046	4	.283	.004
	Post a comment to comfort and show empathy for A	4.155	4	.385	.001
	Post an unrelated comment	4.814	4	.307	.003
	Post a comment to express a similar position as B	5.314	4	.257	.005
	Share B's tweets by retweeting	4.310	4	.366	.001
	Ignore these tweets and continue browsing	13.916	4	.008	.041
	Post a comment condemning B with potentially offensive words	12.539	4	.014	.035

Discussion

Counter-speech is a flexible, bottom-up measure to address hate speech online. However, to date, little is known about the dynamics that counter-speech elicits, specifically, which types of counter-speech facilitate bystander responses that contribute to attenuating the prevalence and dissemination of hate speech. The present study aimed to address this gap in the literature and assessed the effects of four common types of counter-speech on eight bystander behavioral intentions.

Taken together, we showed that three types of counter-speech—namely, abusing the perpetrator, calling on others to intervene, and educating the perpetrator—had no effect on subsequent bystander behavioral intentions, and one strategy—diverting the conversation—had unintended consequences and facilitated bystanders' intentions to ignore hate speech. These findings advance studies that have documented that counter-speech

fosters further bystander engagement in analyses of social media data (Friess et al., 2021; Garland et al., 2022; He et al., 2021; Lasser et al., 2023). Below we discuss all key results starting with the influence of diversionary counter-speech.

It could be speculated that the distraction attempt did divert attention from hate speech to a new topic and, thus, promoted inaction (Barberá et al., 2022; King et al., 2017). However, inaction in response to diversionary counter-speech might also reflect a type of bystander effect (Latané & Darley, 1970), that is, individuals' willingness to intervene in an incident is reduced if other bystanders are present (Darley & Latané, 1968). Notably, the stimulus of diversionary counter-speech presented in the experiment could have been interpreted such that other bystanders (i.e., counter-speaker—user "C") do not consider the incident severe enough to take any targeted action; instead, they bring up an unrelated point. Thus, participants may also not have judged the incident as an emergency that requires a response. Other bystander interventions (e.g., educational counter-speech) did not affect intentions of inaction in the same way, suggesting that diversionary counter-speech offers distinct signals that inform others' reactions. To investigate this explanation, future research should capture participants' perceived severity of the hate speech incident and their sense of responsibility to get engaged. Having said this, inaction can stem from various reasons, such as silent approval of online hate, indifference, fear of retaliation, and a lack of response knowledge (DeSmet et al., 2016; Hansen et al., 2023; Jeyagobi et al., 2022; Song & Oh, 2018). Thus, although response bias is a limitation, incorporating an open-response form in future experimental designs or conducting follow-up interviews could provide valuable insights into the drivers of bystanders' inaction over and beyond counter-speech.

Regarding the disparity between the nil effects we identified and previous studies showing "facilitating" effects of counter-speech, we believe that this could be attributed to differences in the study design. Specifically, studies that found significant effects of counter-speech on bystander reactions often relied on social media data, where large sample sizes make it easier to detect small effect sizes (Friess et al., 2021; Garland et al., 2022; He et al., 2021; Lasser et al., 2023). Our sensitivity analysis highlighted that we were only able to detect medium effects. It is also important to note that previous studies were able to capture long-term effects of counter-speech, taking into account bystander reactions that emerged over a longer period. Our study examined immediate bystander behavioral intentions. It is conceivable that long-term effects could emerge as users gradually internalize the messages conveyed by counter-speech, leading to a more profound change in attitudes and behavior over time. Mechanisms at play might include increased normalization of counter-speech as a social norm, improved collective understanding, and shifts in community standards (Friess et al., 2021; Kunst et al., 2021; Lasser et al., 2023; Wittenbaum et al., 1999). For instance, consistent exposure to counter-speech can gradually influence bystanders' perceptions of acceptable behavior, leading them to more actively support or engage in counter-speech in the future. Additionally, repeated positive reinforcement from counter-speech interactions can build a supportive environment that encourages more constructive bystander reactions over time. These delayed effects may not be captured in immediate behavioral intentions but can significantly impact long-term changes in bystander behavior.

Limitations

The previous conclusions must be considered in light of the following limitations. First, this study only investigated four types of counter-speech in response to one type of hate speech (i.e., racist speech). Different types of online hate, such as misogynistic and anti-LGTBQ+ hate speech, might lead to different bystander responses. More precisely, as participants have likely seen racist hate speech frequently, the overall willingness to react might be underestimated in this study (Fischer et al., 2011; Home Office, 2018; Ortiz, 2019; Soral et al., 2018). Soral and colleagues (2018) emphasized that frequent exposure to hate speech elicits a process of desensitization, which could manifest in downplaying the severity of online hate. When incidents are perceived as less severe, bystander intervention is less likely (Fischer et al., 2011). In addition, the level of severity of an incident might also prompt distinct intervention methods. A study of adolescent bullying found that when cyberbullying incidents were perceived as serious, bystanders were more likely to help by talking to friends/teachers/parents rather than confronting the perpetrator online (Patterson et al., 2017).

Relatedly, as we implemented only one form of hate speech stimulus, results should not be generalized to other incidents without conceptual replication. Specifically, the experimental stimulus included derogatory language but did not call for outgroup violence. The latter may be perceived as more severe and could evoke different bystander

behavioral intentions, regardless of the type of counter-speech (Fischer et al., 2011). To clarify this possibility, future studies ought to explicitly manipulate the severity of the hate speech stimuli.

Furthermore, as in all experiments, the study lacks ecological validity, and participants knew they would not actually perform the chosen behaviors, which could also have led to a nil effect of different types of counter-speech. One way to reduce this concern would be to work in simulated social media platform environments, such as the Mock Social Media Website Tool (Jagayat et al., 2021), where participants can (also) engage with known features of social media platforms to take actions, such as contacting victims or reporting hate speech. Additionally, analyzing observational social media data could clarify dynamics of real-world hate speech conversations. However, it must be noted that the internal validity of these analyses is lower and many crucial bystander reactions are absent in the latter data, especially inaction, reporting, and private messaging with the victim or perpetrator.

Studying online interactions more generally and the impact of counter-speech in particular is inherently complex, as numerous potential influencing factors ought to be considered. As defined, counter-speech is always a direct reply to hate speech. In this study's experimental stimuli, hate speech appeared as a reply, emphasizing the presence of a direct victim (user "A"). On social media, however, many posts lack direct victims (e.g., generalized hate speech like "I hate Chinese people"; ElSherief et al., 2018). Direct hate replies may elicit stronger emotional engagement from bystanders compared to generalized hate speech. Specifically, when people recognize that a direct victim is being attacked, they may interpret the situation as more urgent and severe, such that stronger empathy is evoked. Future experiments using generalized hate speech instead of direct hate replies would provide valuable comparisons to our findings.

Additionally, the observed reactions of others to counter-speech (e.g., "likes/dislikes") are important. It has been well documented that the number of "likes" on content impact attitudes and behavior in various domains (Tiggemann et al., 2018; Zell & Moeller, 2018). More "likes" on the respective posts could indicate greater support for counter-speech, which may enhance its persuasive effect and influence bystanders to act similarly. This moderating factor could be tested by varying the number of "likes" on a counter-speech post and observing differences in bystander reactions. Furthermore, the quantity of comments from others on either counter-speech or hate speech posts may influence other users' reactions to the counter-speech (Friess et al., 2021; Garland et al., 2022; Obermaier et al., 2016; Waddell & Bailey, 2017). For instance, if a counter-speech post receives multiple supportive comments, it may encourage other users to align with the counter-speech. However, if the majority of comments are hostile or mocking, it could undermine the counter-speech and lead bystanders to side with the hate speech instead.

A user's status—whether they are a verified user, influencer, or group leader/member—may affect conversation dynamics as well (Friess et al., 2021; Garland et al., 2022; Jia & Schumann, 2024; Leung et al., 2022). Individuals with greater influence typically have more followers and could gain more attention online, potentially increasing the number of bystanders' reactions and support. Relatedly, when bystanders and victims/perpetrators are connected online or have other personal offline connections, they may be more likely to act in support of either party (High & Buehler, 2019; Jia et al., in preparation). In this study, participants had no personal connection to the perpetrator, victim, or the bystander posting counter-speech; this reflects an incidental exposure to hate speech online. The lack of connection could have reduced participants' willingness to engage overall, suggesting perhaps that without personal connection, counter-speech, regardless of type, fails to mobilize.

Lastly, the effectiveness of counter-speech can also depend on platform characteristics. Different social media platforms vary in user demographics, cultural norms, and technological features, which can significantly affect how counter-speech is perceived and acted on. For instance, some bystander reactions, such as reporting, depend on platform tools like "report" buttons, which may not be consistently available. Similarly, the presence of private messaging features can influence bystander reactions, as individuals might feel more comfortable offering support or intervening privately rather than publicly. Platforms lacking these tools could limit specific reactions like reporting or offering comfort to the victim, making some counter-speech approaches less effective. Additionally, the norms within a platform's community play a crucial role. Empathetic social norms can foster more constructive bystander behaviors, such as reporting and helping victims (Freis & Gurung, 2013; Kunst et al., 2021; Van Cleemput et al., 2014). On platforms where these norms are absent or weaker, such as 4chan, 8kun, Gab, and certain Telegram groups or channels, counter-speech may not have the intended effect and could even provoke hostility. Our participants, however, were not operating in a setting where specific platform norms were enforced. Therefore, the generalizability of our findings across different online spaces is limited. Future research should

explore how features like reporting tools and private messaging, along with platform/community-specific norms, shape the effectiveness of counter-speech in encouraging positive bystander reactions.

Conclusion

Despite these challenges, we believe our study makes an important contribution to the literature. We showed that several types of counter-speech did not influence subsequent bystander behavioral intentions. Additionally, we found that diverting the conversation could evoke inaction, which neither mitigates the spread nor addresses the negative impact of hate speech. Identifying when and why particular modes of counter-speech fail to foster prosocial bystander behavioral intentions is a fruitful avenue for future research.

Footnotes

[1] To ensure the accuracy of the counter-speech stimuli, we also conducted a pilot study (N = 20; see Appendix).

[2] The data used for the *t*-tests did not follow a normal distribution, and the assumption of homogeneity of variance was violated in some tests. It is generally accepted that the *t*-test remains robust when assumptions of normal distribution are not fully met. To address concerns about the lack of homogeneity of variances, we implemented Welch *t*-tests (see the Supplementary Material on the associated OSF repository here for details on assumption testing).

[3] The data used for analysis violated multiple assumptions of running ANOVA/MANOVA (see the Supplementary Material on the associated OSF repository here for details on assumption testing). Under the violations of assumptions, the Kruskal-Wallis test is more powerful at detecting differences among treatments than the ANOVA *F*-test. However, due to G*Power's limitations in conducting sensitivity analysis for the Kruskal-Wallis test directly, one-way ANOVA was used as an alternative approach in the sensitivity analysis.

Conflict of Interest

The authors have no conflicts of interest to declare.

Use of AI Services

The authors declare they have used AI services, specifically Grammarly, for grammar correction and minor style refinements. They carefully reviewed all suggestions from these services to ensure the original meaning and factual accuracy were preserved.

Authors' Contribution

Yue Jia: writing—original draft, writing—review & editing, conceptualization, methodology, investigation, formal analysis, visualization. **Sandy Schumann:** writing—review & editing, supervision, methodology, investigation, funding acquisition.

The first author, Yue Jia, passed away shortly before a minor second round of revisions and the proofs of the article could be completed. The second author, Sandy Schumann, completed these revisions and the proofs.

References

Balkin, J. M. (2017). Free speech in the algorithmic society: Big data, private governance, and new school speech regulation. *UC Davis Law Review*, *51*, 1149–1210. https://heinonline.org/HOL/LandingPage?handle=hein.journals/davlr51&div=40&id=&page=

Barberá, P., Gohdes, A. R., Iakhnis, E., & Zeitzoff, T. (2022). Distract and divert: How world leaders use social media during contentious politics. *The International Journal of Press/Politics, 29*(1), 47–73. https://doi.org/10.1177/19401612221102030

Barlińska, J., Szuster, A., & Winiewski, M. (2013). Cyberbullying among adolescent bystanders: Role of the communication medium, form of violence, and empathy. *Journal of Community & Applied Social Psychology*, *23*(1), 37–51. https://doi.org/10.1002/casp.2137

Bartlett, J., & Krasodomski-Jones, A. (2015). *Counter-speech examining content that challenges extremism online*. DEMOS. https://demos.co.uk/wp-content/uploads/2015/10/Counter-speech-1.pdf

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

Benesch, S., Ruths, D., Dillon, K. P., Saleem, H. M., & Wright, L. (2016, October 15). *Considerations for successful counterspeech*. Dangerous speech project. https://dangerousspeech.org/considerations-for-successful-counterspeech/

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

Bergmann, M. C., & Baier, D. (2018). Prevalence and correlates of cyberbullying perpetration. Findings from a German representative student survey. *International Journal of Environmental Research and Public Health*, *15*(2), Article 274. https://doi.org/10.3390/ijerph15020274

Brown, A. (2018). What is so special about online (as compared to offline) hate speech? *Ethnicities*, *18*(3), 297–326. https://doi.org/10.1177/1468796817709846

Buerger, C. (2022). Counterspeech: A literature review. *SSRN*. https://dx.doi.org/10.2139/ssrn.4066882

Bundesministerium der Justiz. (2017, September 7). *Network Enforcement Act (NetzDG)*. https://www.bmj.de/SharedDocs/Gesetzgebungsverfahren/DE/2017_NetzDG.html?nn=17134

Cambridge Dictionary. (n.d.). *Hate speech*. https://dictionary.cambridge.org/us/dictionary/english/hate-speech

Cepollaro, B., Lepoutre, M., & Simpson, R. M. (2023). Counterspeech. *Philosophy Compass*, *18*(1), Article e12890. https://doi.org/10.1111/phc3.12890

Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2017). Anyone can become a troll: Causes of trolling behavior in online discussions. In *CSCW '17: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 1217–1230). Association for Computing Machinery. https://doi.org/10.1145/2998181.2998213

Chetty, N., & Alathur, S. (2018). Hate speech review in the context of online social networks. *Aggression and Violent Behavior*, *40*, 108–118. https://doi.org/10.1016/j.avb.2018.05.003

Copland, S. (2020). Reddit quarantined: Can changing platform affordances reduce hateful material online? *Internet Policy Review*, *9*(4), 1–26. http://hdl.handle.net/10419/225653

Corazza, M., Menini, S., Cabrio, E., Tonelli, S., & Villata, S. (2020). A multilingual evaluation for online hate speech detection. *ACM Transactions on Internet Technology*, *20*(2), Article 10. https://doi.org/10.1145/3377323

Council of Europe. (n.d.). *Hate speech*. https://www.coe.int/en/web/freedom-expression/hate-speech

Crawford, K., & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society*, *18*(3), 410–428. https://doi.org/10.1177/1461444814543163

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

De Gibert, O., Perez, N., García-Pablos, A., & Cuadros, M. (2018). Hate speech dataset from a white supremacy forum. *arXiv*. https://doi.org/10.48550/arXiv.1809.04444

DeSmet, A., Bastiaensens, S., Van Cleemput, K., Poels, K., Vandebosch, H., Cardon, G., & De Bourdeaudhuij, I. (2016). Deciding whether to look after them, to like it, or leave it: A multidimensional analysis of predictors of positive and negative bystander behavior in cyberbullying among adolescents. *Computers in Human Behavior*, *57*, 398–415. https://doi.org/10.1016/j.chb.2015.12.051

DeSmet, A., Veldeman, C., Poels, K., Bastiaensens, S., Van Cleemput, K., Vandebosch, H., & De Bourdeaudhuij, I. (2014). Determinants of self-reported bystander behavior in cyberbullying incidents amongst adolescents. *Cyberpsychology, Behavior, and Social Networking*, *17*(4), 207–215. https://doi.org/10.1089/cyber.2013.0027

Edwards, S. M., Li, H., & Lee, J.-H. (2002). Forced exposure and psychological reactance: Antecedents and consequences of the perceived intrusiveness of pop-up ads. *Journal of Advertising*, *31*(3), 83–95. https://doi.org/10.1080/00913367.2002.10673678

ElSherief, M., Kulkarni, V., Nguyen, D., Wang, W. Y., & Belding, E. (2018). Hate lingo: A target-based linguistic analysis of hate speech in social media. *Proceedings of the International AAAI Conference on Web and Social Media, 12*(1), 42–51. https://doi.org/10.1609/icwsm.v12i1.15041

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

Farid, H. (2021). An overview of perceptual hashing. *Journal of Online Trust and Safety*, *1*(1), 1–22. https://doi.org/10.54501/jots.v1i1.24

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, *59*(7), 96–104. https://doi.org/10.1145/2818717

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

Freis, S. D., & Gurung, R. A. (2013). A Facebook analysis of helping behavior in online bullying. *Psychology of Popular Media Culture*, *2*(1), 11–19. https://psycnet.apa.org/doi/10.1037/a0030239

Friess, D., Ziegele, M., & Heinbach, D. (2021). Collective civic moderation for deliberation? Exploring the links between citizens' organized engagement in comment sections and the deliberative quality of online discussions. *Political Communication*, *38*(5), 624–646. https://doi.org/10.1080/10584609.2020.1830322

Garland, J., Ghazi-Zahedi, K., Young, J. G., Hébert-Dufresne, L., & Galesic, M. (2020). Countering hate on social media: Large scale classification of hate and counter speech. *arXiv*. https://doi.org/10.48550/arXiv.2006.01974

Garland, J., Ghazi-Zahedi, K., Young, J. G., Hébert-Dufresne, L., & Galesic, M. (2022). Impact and dynamics of hate and counter speech online. *EPJ Data Science*, *11*(1), Article 3. https://doi.org/10.1140/epjds/s13688-021-00314-6

Griffin, R. (2022). New school speech regulation as a regulatory strategy against hate speech on social media: The case of Germany's NetzDG. *Telecommunications Policy*, *46*(9), Article 102411. https://doi.org/10.1016/j.telpol.2022.102411

Han, S.-H., & Brazeal, L. M. (2015). Playing nice: Modeling civility in online political discussions. *Communication Research Reports*, *32*(1), 20–28. https://doi.org/10.1080/08824096.2014.989971

Han, S.-H., Brazeal, L. M., & Pennington, N. (2018). Is civility contagious? Examining the impact of modeling in online political discussions. *Social Media+Society*, *4*(3), 1–12. https://doi.org/10.1177/2056305118793404

Hangartner, D., Gennaro, G., Alasiri, S., Bahrich, N., Bornhoft, A., Boucher, J., Demirci, B. B., Derksen, L., Hall, A., Jochum, M., Munoz, M. M., Richter, M., Vogel, F., Wittwer, S., Wüthrich, F., Gilardi, F., & Donnay, K. (2021). Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences*, *118*(50), Article e2116310118. https://doi.org/10.1073/pnas.2116310118

Hansen, T. M., Lindekilde, L., & Karg, S. T. S. (2023). The devil is in the detail: Reconceptualising bystander reactions to online political hostility. *Behaviour & Information Technology, 43*(14), 3523–3536. https://doi.org/10.1080/0144929X.2023.2282653

He, B., Ziems, C., Soni, S., Ramakrishnan, N., Yang, D., & Kumar, S. (2021). Racism is a virus: Anti-Asian hate and counterspeech in social media during the COVID-19 crisis. In *ASONAM '21: Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 90–94). Association for Computing Machinery. https://doi.org/10.1145/3487351.3488324

High, A. C., & Buehler, E. M. (2019). Receiving supportive communication from Facebook friends: A model of social ties and supportive communication in social network sites. *Journal of Social and Personal Relationships*, *36*(3), 719–740. https://doi.org/10.1177/0265407517742978

Hoffman, M. L. (2014). Empathy, social cognition, and moral action. In W. M. Kurtines, J. Gewirtz, & J. L. Lamb (Eds.), *Handbook of moral behavior and development* (pp. 299–326). Psychology Press. https://www.taylorfrancis.com/chapters/edit/10.4324/9781315807294-15/empathy-social-cognition-moral-action-martin-hoffman

Home Office. (2018, October 16). *Hate crime, England and Wales, 2017 to 2018*. GOV.UK. https://www.gov.uk/government/statistics/hate-crime-england-and-wales-2017-to-2018

Howard, J. W. (2021). Terror, hate and the demands of counter-speech. *British Journal of Political Science*, *51*(3), 924–939. https://doi.org/10.1017/S000712341900053X

Human Rights Watch. (2018, February 14). *Germany: Flawed social media law*. https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law

Jagayat, A., Boparai, G., Pun, C., & Choma, B. L. (2021). *Mock Social Media Website Tool (1.0)* [Computer software]. Mock Social Media Website Tool. https://docs.studysocial.media

Jeyagobi, S., Munusamy, S., Kamaluddin, M. R., Ahmad Badayai, A. R., & Kumar, J. (2022). Factors influencing negative cyber-bystander behavior: A systematic literature review. *Frontiers in Public Health*, *10*, Article 965017. https://doi.org/10.3389/fpubh.2022.965017

Jhaver, S., Boylston, C., Yang, D., & Bruckman, A. (2021). Evaluating the effectiveness of deplatforming as a moderation strategy on Twitter. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), Article 381. https://doi.org/10.1145/3479525

Jia, Y., & Schumann, S. (2024). *Examining dynamics of counter-speech to hate speech on X.* [Manuscript in preparation]. Social Research Institute, University College London.

Kennedy, C. J., Bacon, G., Sahn, A., & von Vacano, C. (2020). Constructing interval variables via faceted Rasch measurement and multitask deep learning: A hate speech application. *arXiv*. https://doi.org/10.48550/arXiv.2009.10277

Kim, Y., Baek, T. H., Yoon, S., Oh, S., & Choi, Y. K. (2017). Assertive environmental advertising and reactance: Differences between South Koreans and Americans. *Journal of Advertising*, *46*(4), 550–564. https://doi.org/10.1080/00913367.2017.1361878

King, G., Pan, J., & Roberts, M. E. (2017). How the Chinese government fabricates social media posts for strategic distraction, not engaged argument. *American Political Science Review*, *111*(3), 484–501. https://doi.org/10.1017/S0003055417000144

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(24), 8788–8790. https://doi.org/10.1073%2Fpnas.1320040111

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: A literature review. *Quality & Quantity*, *47*(4), 2025–2047. https://doi.org/10.1007/s11135-011-9640-9

Kunst, M., Porten-Cheé, P., Emmer, M., & Eilders, C. (2021). Do "good citizens" fight hate speech online? Effects of solidarity citizenship norms on user responses to hate comments. *Journal of Information Technology & Politics*, *18*(3), 258–273. https://doi.org/10.1080/19331681.2020.1871149

Lasser, J., Herderich, A., Garland, J., Aroyehun, S. T., Garcia, D., & Galesic, M. (2023). Collective moderation of hate, toxicity, and extremity in online discussions. *arXiv*. https://doi.org/10.48550/arXiv.2303.00357

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

Leung, F. F., Gu, F. F., & Palmatier, R. W. (2022). Online influencer marketing. *Journal of the Academy of Marketing Science*, *50*(2), 226–251. https://doi.org/10.1007/s11747-021-00829-4

Maarouf, A., Pröllochs, N., & Feuerriegel, S. (2022). The virality of hate speech on social media. *arXiv*. https://doi.org/10.48550/arXiv.2210.13770

Macaulay, P. J. R., Betts, L. R., Stiller, J., & Kellezi, B. (2022). Bystander responses to cyberbullying: The role of perceived severity, publicity, anonymity, type of cyberbullying, and victim response. *Computers in Human Behavior*, *131*, Article 107238. https://doi.org/10.1016/j.chb.2022.107238

Machackova, H., Dedkova, L., & Mezulanikova, K. (2015). Brief report: The bystander effect in cyberbullying incidents. *Journal of Adolescence*, *43*(1), 96–99. https://doi.org/10.1016/j.adolescence.2015.05.010

Mathew, B., Kumar, N., Goyal, P., & Mukherjee, A. (2018). Analyzing the hate and counter speech accounts on Twitter. *arXiv*. https://doi.org/10.48550/arXiv.1812.02712

Mathew, B., Saha, P., Tharad, H., Rajgaria, S., Singhania, P., Maity, S. K., Goyal, P., & Mukherjee, A. (2019). Thou shalt not hate: Countering online hate speech. *Proceedings of the International AAAI Conference on Web and Social Media*, *13*(01), 369–380. https://doi.org/10.1609/icwsm.v13i01.3237

McDevitt, J., Levin, J., & Bennett, S. (2002). Hate crime offenders: An expanded typology. *Journal of Social Issues*, *58*(2), 303–317. https://doi.org/10.1111/1540-4560.00262

Media Smarts. (n.d.). *Online hate*. https://mediasmarts.ca/digital-media-literacy/digital-issues/online-hate

Miron, A. M., & Brehm, J. W. (2006). Reactance theory - 40 years later. *Zeitschrift für Sozialpsychologie*, *37*(1), 9–18. https://doi.org/10.1024/0044-3514.37.1.9

Miškolci, J., Kováčová, L., & Rigová, E. (2020). Countering hate speech on Facebook: The case of the Roma minority in Slovakia. *Social Science Computer Review*, *38*(2), 128–146. https://doi.org/10.1177/0894439318791786

Molina, R. G., & Jennings, F. J. (2018). The role of civility and metacommunication in Facebook discussions. *Communication Studies*, *69*(1), 42–66. https://doi.org/10.1080/10510974.2017.1397038

Mondal, M., Silva, L. A., Correa, D., & Benevenuto, F. (2018). Characterizing usage of explicit hate expressions in social media. *New Review of Hypermedia and Multimedia*, *24*(2), 110–130. https://doi.org/10.1080/13614568.2018.1489001

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

Obermaier, M., Schmuck, D., & Saleem, M. (2021). I'll be there for you? Effects of Islamophobic online hate speech and counter speech on Muslim in-group bystanders' intention to intervene. *New Media & Society, 25*(9), 2339–2358. https://doi.org/10.1177/14614448211017527

Oksanen, A., Hawdon, J., Holkeri, E., Näsi, M., & Räsänen, P. (2014). Exposure to online hate among young social media users. In M. N. Warehime (Ed.), *Soul of society: A focus on the lives of children & youth* (pp. 253–273). Emerald. https://www.emerald.com/insight/content/doi/10.1108/s1537-466120140000018021/full/html

Ortiz, S. M. (2019). "You can say I got desensitized to it": How men of color cope with everyday racism in online gaming. *Sociological Perspectives*, *62*(4), 572–588. https://doi.org/10.1177/0731121419837588

Pacheco, E., & Melhuish, N. (2018). *Online hate speech: A survey on personal experiences and exposure among adult New Zealanders*. Netsafe. https://dx.doi.org/10.2139/ssrn.3272148

Patterson, L. J., Allan, A., & Cross, D. (2017). Adolescent bystander behavior in the school and online environments and the implications for interventions targeting cyberbullying. *Journal of School Violence*, *16*(4), 361–375. https://doi.org/10.1080/15388220.2016.1143835

Perugini, M., Gallucci, M., & Costantini, G. (2018). A practical primer to power analysis for simple experimental designs. *International Review of Social Psychology*, *31*(1), Article 20. https://dx.doi.org/10.5334/IRSP.181

Quick, B. L., & Stephenson, M. T. (2007). Further evidence that psychological reactance can be modeled as a combination of anger and negative cognitions. *Communication Research*, *34*(3), 255–276. https://doi.org/10.1177/0093650207300427

Räsänen, P., Hawdon, J., Holkeri, E., Keipi, T., Näsi, M., & Oksanen, A. (2016). Targets of online hate: Examining determinants of victimization among young Finnish Facebook users. *Violence and Victims*, *31*(4), 708–725. https://doi.org/10.1891/0886-6708.VV-D-14-00079

Rieger, D., Kümpel, A. S., Wich, M., Kiening, T., & Groh, G. (2021). Assessing the extent and types of hate speech in fringe communities: A case study of alt-right communities on 8chan, 4chan, and Reddit. *Social Media+Society*. https://journals.sagepub.com/doi/full/10.1177/20563051211052906

Rogers, R. (2020). Deplatforming: Following extreme internet celebrities to Telegram and alternative social media. *European Journal of Communication*, *35*(3), 213–229. https://doi.org/10.1177/0267323120922066

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

Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., & Pierrehumbert, J. B. (2021). HateCheck: Functional tests for hate speech detection models. *arXiv*. https://doi.org/10.48550/arXiv.2012.15606

Song, J., & Oh, I. (2018). Factors influencing bystanders' behavioral reactions in cyberbullying situations. *Computers in Human Behavior*, *78*, 273–282. https://doi.org/10.1016/j.chb.2017.10.008

Soral, W., Bilewicz, M., & Winiewski, M. (2018). Exposure to hate speech increases prejudice through desensitization. *Aggressive Behavior*, *44*(2), 136–146. https://doi.org/10.1002/ab.21737

Steindl, C., Jonas, E., Sittenthaler, S., Traut-Mattausch, E., & Greenberg, J. (2015). Understanding psychological reactance: New developments and findings. *Zeitschrift für Psychologie*, *223*(4), 205–214. https://doi.org/10.1027/2151-2604/a000222

Stop Hate UK. (n.d.). *Online hate and free speech*. https://www.stophateuk.org/about-hate-crime/what-is-online-hate-crime/online-hate-and-free-speech/

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

Thomas, L., Falconer, S., Cross, D., Monks, H., & Brown, D. (2012). *Cyberbullying and the bystander*. Australian Human Rights Commission. https://humanrights.gov.au/our-work/childrens-rights/publications/cyberbullying-and-bystander

Tiggemann, M., Hayden, S., Brown, Z., & Veldhuis, J. (2018). The effect of Instagram "likes" on women's social comparison and body dissatisfaction. *Body Image*, *26*, 90–97. https://doi.org/10.1016/j.bodyim.2018.07.002

Trifiro, B. M., & Gerson, J. (2019). Social media usage patterns: Research note regarding the lack of universal validated measures for active and passive use. *Social Media+Society*. https://doi.org/10.1177/2056305119848743

Tsugawa, S., & Ohsaki, H. (2015). Negative messages spread rapidly and widely on social media. In *COSN '15: Proceedings of the 2015 ACM on Conference on Online Social Networks* (pp. 151–160). Association for Computing Machinery. https://doi.org/10.1145/2817946.2817962

UK Legislation. (n.d.). *Public Order Act 1986*. https://www.legislation.gov.uk/ukpga/1986/64

UK Legislation. (2023). *Online Safety Act 2023*. https://www.legislation.gov.uk/ukpga/2023/50/enacted

Ullmann, S., & Tomalin, M. (2020). Quarantining online hate speech: Technical and ethical perspectives. *Ethics and Information Technology*, *22*, 69–80. https://doi.org/10.1007/s10676-019-09516-z

Uyheng, J., & Carley, K. M. (2020). Bots and online hate during the COVID-19 pandemic: Case studies in the United States and the Philippines. *Journal of Computational Social Science*, *3*(2), 445–468. https://doi.org/10.1007/s42001-020-00087-4

Van Cleemput, K., Vandebosch, H., & Pabian, S. (2014). Personal characteristics and contextual factors that determine "helping," "joining in," and "doing nothing" when witnessing cyberbullying. *Aggressive Behavior*, *40*(5), 383–396. https://doi.org/10.1002/ab.21534

Veilleux-Lepage, Y. (2016). Retweeting the caliphate. The role of soft-sympathizers in the Islamic State's social media strategy. *Turkish Journal of Security Studies*, *18*(1), 53–69. https://www.researchgate.net/publication/273896091_Retweeting_the_Caliphate_The_Role_of_Soft-Sympathizers_in_the_Islamic_State%27s_Social_Media_Strategy

Vidgen, B., Botelho, A., Broniatowski, D., Guest, E., Hall, M., Margetts, H., Tromble, R., Waseem, Z., & Hale, S. (2020). Detecting East Asian prejudice on social media. *arXiv*. https://doi.org/10.48550/arXiv.2005.03909

Vidgen, B., Margetts, H., & Harris, A. (2019). *How much online abuse is there*. The Alan Turing Institute. https://www.turing.ac.uk/sites/default/files/2019-11/online_abuse_prevalence_full_24.11.2019_-_formatted_0.pdf

Vidgen, B., Thrush, T., Waseem, Z., & Kiela, D. (2021). Learning from the worst: Dynamically generated datasets to improve online hate detection. *arXiv*. https://doi.org/10.48550/arXiv.2012.15761

Waddell, T. F., & Bailey, A. (2017). Inspired by the crowd: The effect of online comments on elevation and universal orientation. *Communication Monographs*, *84*(4), 534–550. https://doi.org/10.1080/03637751.2017.1369137

Wilson, J. Q., & Kelling, G. L. (1982). Broken windows. *Atlantic Monthly*, *249*(3), 29–38. http://faculty.washington.edu/matsueda/courses/587/readings/Wilson%20and%20Kelling%20(1982).pdf

Winiewski, M., Hansen, K., Bilewicz, M., Soral, W., Świderska, A., & Bulska, D. (2017). *Contempt speech, hate speech*. Stefan Batory Foundation. https://www.ngofund.org.pl/wp-content/uploads/2017/02/Contempt_Speech_Hate_Speech_Full_Report.pdf

Wittenbaum, G. M., Hubbell, A. P., & Zuckerman, C. (1999). Mutual enhancement: Toward an understanding of the collective preference for shared information. *Journal of Personality and Social Psychology, 77*(5), 967–978. https://doi.org/10.1037/0022-3514.77.5.967

Zell, A. L., & Moeller, L. (2018). Are you happy for me… on Facebook? The potential importance of "likes" and comments. *Computers in Human Behavior*, *78*, 26–33. https://doi.org/10.1016/j.chb.2017.08.050

Zemack-Rugar, Y., Moore, S. G., & Fitzsimons, G. J. (2017). Just do it! Why committed consumers react negatively to assertive ads. *Journal of Consumer Psychology*, *27*(3), 287–301. https://doi.org/10.1016/j.jcps.2017.01.002

Zhu, H. (2020). Countering COVID-19-related anti-Chinese racism with translanguaged swearing on social media. *Multilingua*, *39*(5), 607–616. https://doi.org/10.1515/multi-2020-0093

Zhu, W., & Bhat, S. (2021). Generate, prune, select: A pipeline for counterspeech generation against online hate speech. *arXiv*. https://doi.org/10.48550/arXiv.2106.01625

Appendix

Pilot Study

This study was conducted during the survey design phase and aimed to ensure that all counter-speech stimuli (original version) were perceived in line with their intended purpose. $N = 20$ participants were recruited, and each read all four types of original counter-speech and then answered four identical questions accordingly: *How strongly do you agree or disagree with the following statements about C's counter-speech?*. Five answers were provided: *C aims to abuse B*; *C aims to educate and persuade B*; *C aims to divert the conversation and distract people*; *C aims to directly call on others to intervene in B's hate speech*; and *C aims to express humor*. Answer options were indicated on a 5-point Likert-type scale (1 – *strongly disagree*, 5 – *strongly agree*). One-sample *t*-tests with the test value of $M > 3$ (thus, representing the average answer options *Agree* or *Strongly agree*) were performed separately in each of the four questions (4 * 5 = 20 in total). As shown in Table A1, all target answers that expressed the intended purpose of the corresponding counter-speech rejected the null hypothesis of $M \leq 3$, and all non-target answers failed to reject the null hypothesis. As a result, all original counter-speech stimuli were perceived in line with their intended purpose. However, the results were not perfect. For example, in question 3, some participants believed that the original counter-speech designed to divert the conversation also aimed to express humor ($M = 3.05$). In question 4, some participants believed that the original counter-speech designed to abuse perpetrators also aimed to divert the conversation ($M = 3.05$) and call on others to intervene ($M = 3.40$). After asking participants for their views, we updated the original version of the four types of counter-speech to the version used in the formal experiment.

Table A1. *Results of the Pilot Study.*

Counter-speech (original version)	Answer (#: target answer; C: bystander; B: perpetrator)	*M*	*SD*	*p* *Ha* ($M > 3$)
Type 1 (...educate...) ($n = 20$)	C aims to abuse B	1.25	0.43	> .999
	C aims to educate and persuade B #	4.50	0.59	< .001*
	C aims to divert the conversation and distract people	2.50	1.20	.961
	C aims to directly call on others to intervene in B's hate speech	2.75	1.22	.815
	C aims to express humor	1.60	0.86	> .999*
Type 2 (...call on...) ($n = 20$)	C aims to abuse B	2.50	1.50	.924
	C aims to educate and persuade B	2.70	1.19	.863
	C aims to divert the conversation and distract people	2.40	1.43	.962
	C aims to directly call on others to intervene in B's hate speech #	4.60	0.58	< .001*
	C aims to express humor	1.20	0.60	> .999
Type 3 (...divert...) ($n = 20$)	C aims to abuse B	1.50	0.87	> .999
	C aims to educate and persuade B	1.45	0.86	> .999
	C aims to divert the conversation and distract people #	4.70	0.46	< .001*
	C aims to directly call on others to intervene in B's hate speech	1.70	0.90	> .999
	C aims to express humor	3.05	1.21	.428
Type 4 (...abuse...) ($n = 20$)	C aims to abuse B #	4.70	0.56	< .001*
	C aims to educate and persuade B	1.65	0.91	> .999
	C aims to divert the conversation and distract people	3.05	1.24	.429
	C aims to directly call on others to intervene in B's hate speech	3.40	1.11	.062
	C aims to express humor	1.30	0.56	> .999

Note. $^*p < .050$.

Supplementary materials

Supplementary materials including Assumption Test Result, Survey Instrument and Results of the MANCOVA are available in the associated OSF repository here.

About Authors

Yue Jia (MSc, UCL) was a doctoral student at UCL's Social Research Institute. His research focused on dynamics and implications of counter-speech to hate speech online.

Sandy Schumann (PhD, Université Libre de Bruxelles) is a Lecturer (Assistant Professor) at University College London, Department of Security and Crime Science. Her current research examines questions that pertain to the interplay between technology and hate (crime or speech), extremism, as well as radicalization. She takes an interdisciplinary approach and works in the borderlands of communication science, social psychology, and terrorism studies. Dr Schumann is keen to translate her research into policy and practice. Doing so, she works with governments and civil society organizations to conduct impact evaluations.

https://orcid.org/0000-0002-0900-5356

Correspondence to
Sandy Schumann, Department of Security and Crime Science, University College London, 35 Tavistock Square, London WC1H 9EZ, UK, s.schumann@ucl.ac.uk