



ARTICLE



<https://doi.org/10.1057/s41599-024-02761-8>

OPEN

Bystanders' collective responses set the norm against hate speech

Jimena Zapata^{1,2}[✉], Justin Sulik³, Clemens von Wulffen⁴ & Ophelia Deroy^{1,5,6}

Hate speech incidents often occur in social settings, from public transport to football stadiums. To counteract a prevailing passive attitude towards them, governmental authorities, sociologists, and philosophers stress bystanders' responsibility to oppose or block hate speech. Here, across two online experiments with UK participants using custom visual vignettes, we provide empirical evidence that bystanders' expression of opposition can affect how harmful these incidents are perceived, but only as part of a collective response: one expressed by a majority of bystanders present. Experiment 1 ($N = 329$) shows that the silence or intervention of three bystanders affects the harm caused by hate speech, but one bystander does not. Experiment 2 ($N = 269$) shows this is not simply a matter of numbers but rather one of norms: only unanimous opposition reduces the public perception of the damage created by the incident. Based on our results, we advance an empirical norm account: group responses to hate speech modulate its harm by indicating either a permissive or a disapproving social norm. Our account and results, showing the need to consider responses to hate speech at a collective level, have direct implications for social psychology, the philosophy of language and public policies.

¹Faculty of Philosophy, Philosophy of Science and Religious Studies Ludwig-Maximilians-Universität München, Munich, Germany. ²Department of Philosophy I, University of Granada, Granada, Spain. ³Cognition, Values & Behaviour Lab, Faculty of Philosophy, Philosophy of Science and Religious Studies Ludwig-Maximilians-Universität München, Munich, Germany. ⁴Department of Experimental Psychology, University of Oxford, Oxford, UK. ⁵Munich Center for Neuroscience, Munich, Germany. ⁶Institute of Philosophy, School of Advanced Study, University of London, London, UK. ✉email: jzapata@lmu.de

Introduction

Is remaining silent when witnessing a hate speech attack harmful? Conversely, does speaking out against hate speech reduce the harm the attack creates? Given that this demeaning speech is harmful (Maitra and McGowan, 2012; Waldron, 2012; Walters, 2014; Zapata and Deroy, 2023), most theoretical approaches to hate speech argue that silent bystanders could unintentionally support the aggressors (Langton, 2007, 2012, 2018a, 2018b; Maitra, 2004). This support could consist in letting perpetrators informally gain practical authority to express hateful derogatory statements (Langton, 2018a; Witek, 2013), normalising the verbal abuse of targeted victims (Ayala and Vasilyeva, 2016) and creating more stress and suffering for them and, by extension, society (Gelber and McNamara, 2016; Goldberg, 2010; 2020; Janson et al., 2009).

Recent investigations emphasise that the norm against hate may be unpopular for a population segment and that compliance is contingent on the belief that the majority agrees (Álvarez-Benjumea, 2023). Furthermore, when people more likely to hold xenophobic attitudes are exposed to hate speech, the tendency to imitate such behaviour increases (Bicchieri, 2005; Diekmann et al., 2015; Bicchieri and Dimant, 2022). Therefore, signalling a robust social norm against hate (i.e., voicing opposition) may be crucial in effectively restraining the rise of hate speech (Álvarez-Benjumea and Winter, 2020).

Under this assumption, intensive research has explored the impact of actively responding by showing opposition to hate speech (Álvarez-Benjumea, 2023; Gelber, 2012; Howard, 2021; Lepoutre, 2017; 2019; de Silva and Simpson, 2022); analysing contextual determinants (e.g., speaker's identity features; the cultural background of the group; the number of bystanders present or how dangerous the incident is perceived to be) that favour or disfavour bystanders' intervention (Hornsey and Imani, 2004; Dessel et al., 2017; Dickter and Newton, 2013; Gulker et al., 2013; Gibson et al., 2020; Rovira et al., 2021; Wong et al., 2021; Leonhard et al., 2018); investigating the best practices on how (e.g., by shaming, correcting or taking distance from perpetrators) and when to oppose hateful remarks (Fumagalli, 2021; Gagliardone et al., 2015, Lepoutre, 2017); and identifying which subjects (e.g., those perceived as higher in hierarchy) are better placed to respond to hate speech (Ashburn-Nardo et al., 2014; 2020). Researchers have also shown that getting involved in counter speech may be extremely challenging and costly for individual bystanders and even more so for targets of hate speech, who are the actual victims of those attacks (Czopp, Monteith (2003); Nielsen, 2012; Dickter and Newton, 2013; Langton, 2018b; McGowan, 2018).

Yet the initial questions remain untested: Do ordinary citizens perceive hate speech incidents as more harmful when they occur in front of silent, passive bystanders? Do third-party observers consider bystanders who voice their opposition helpful in reducing the harm created by hate speech incidents? Our study addresses these two core questions, exploring under which circumstances and contexts ordinary people perceive hate speech as harmful and an opposing response as effective in reducing the harm created by such incidents.

It is important to add that we conducted our study with participants from the UK, where reporting and recording hate crimes heavily rely on victims' and bystanders' perceptions (Boushehrian, 2020). The definition of hate crimes used by the police and the Crown Prosecution Service in the UK (CPS) states: "Any criminal offence which is perceived by the victim or any other person, to be motivated by hostility or prejudice, based on a person's disability or perceived disability; race or perceived race; religion or perceived religion; sexual orientation or perceived sexual orientation; transgender identity or perceived transgender

identity", which confirms that studying ordinary people's perceptions of hate incidents is of significant relevance.

Additionally, empirical research has confirmed that witnessing repetitive verbal mistreatment and abusive discrimination affects both bystanders and direct targets in similar physiological and psychological ways (Janson and Hazler, 2004; Janson et al., 2009; Perry and Alvi, 2012). Exposure to hate speech has been linked with more significant desensitisation to demeaning expressions (Greenberg and Pyszczynski, 1985) and a decreasing sympathy for the targets of hate speech, which reinforces outgroup prejudice (Leets, 2001; Soral et al., 2018), eroding social coexistence. Therefore, studying how hate speech functions in the eyes of ordinary citizens may better inform public policies directed to them that aim to change the apparent leniency towards hate speech harm (Cook and Sheppard, 2018), helping to identify the most effective strategies to counter hate speech (Gulker et al., 2013).

A satisfactory response to our research questions should also illuminate why or how bystanders' responses could reduce the harmful effects of hate speech. Here we hypothesised that people perceive the same attack as less harmful when it occurs in a place where showing opposition is the social norm in place. Besides showing that a normative social context significantly shapes individuals' attitudes towards racism (Blanchard et al., 1994; Monteith et al., 1996; Zitek and Hebl 2007), researchers have shown that discrimination and its harms increase if society allows shared norms prohibiting discrimination to be eroded by whatever means (Barr et al., 2018). Then, we find it essential to answer whether people perceive the same attack as less harmful when it occurs in a place where showing opposition is the social norm.

Here, we take social norms to be unwritten rules and regularities that occur in a social context and create shared expectations within a group about how people should behave in certain situations (Bicchieri, 2016; House, 2018). They regulate social interactions in an informal and often subtle way by changing individuals' social expectations (Przepiorka et al., 2022; Opp, 2001). Some examples include tipping at a restaurant, choosing the proper way to greet a stranger, how we talk or eat, but also norms that support unpopular, inequitable, or dysfunctional social outcomes, such as the persistence of the gender pay gap, tolerance of hate speech, or female genital mutilation (Przepiorka et al., 2022). They are temporary and subject to change, as happened with the social rule of not smoking in enclosed spaces (Opp, 2002; Bicchieri and Mercier, 2014).

Studying responses to hate speech through visual vignettes

People's responses to demeaning and offensive language have been analysed mainly using written vignettes (e.g., Swim and Hyers, 1999; de Araujo et al. (2022); Almagro et al., 2022). This type of vignette consists of short, carefully constructed descriptions offering a systematic combination of characteristics of persons, objects or situations. It is widely used in social sciences to investigate respondents' beliefs, attitudes, or judgements (Atzmüller and Steiner, 2010). Their effectiveness has been demonstrated, especially in sensitive research topics such as abuse, trauma, stigma, social injustice, sexuality or mental health, where data quality benefited from participants distancing themselves from personal circumstances when answering surveys or questionnaires (Khanolainen and Semenova, 2020).

However, written vignettes also face problems because they offer scarce contextual information due to their word limit, making it challenging to reflect the richness of real-life situations and contextual determinants crucial to understanding some problematic cases (Parkinson and Manstead, 1993). To address

this, researchers have made use of artistic visual material, demonstrating that offering images in addition to written information allows participants to better understand the situations they evaluate (Holm et al., 2018; Khanolainen and Semenova, 2020), notably in sensitive topics such as bullying or verbal abuse. We followed that line of research and created a battery of cartoons as visually enhanced vignettes for our study.

Using cartoons allowed us to easily show participants many aspects of the incidents that otherwise would require extensive descriptions: specific features of the physical appearance and facial expressions of perpetrators, bystanders and victims; their body language, the physical distance between bystanders and the attack, the public nature of the space where the attack occurs, and most importantly, whether the bystanders present responded individually or collectively, following the majority or against it. For example, we could show the perpetrators' disdain and dislike for the victims or the defencelessness of the racialised victims through their facial expressions, and we could make it clear that all bystanders had the opportunity to react against the attack by locating them close to the incident and by directing their lines of sight to the attack. By including those features, we provided participants with relevant information about the incident's social context and, at the same time, reduced the scope of subjective interpretations, making the experiment less demanding and allowing participants to focus on the questions presented.

For example, researchers have shown that derogatory language is perceived as more or less permissible depending on whether it is used by someone who shares group membership with the target (Henry et al., 2014; Almagro et al., 2022). Then, by standardising the appearance of perpetrators and bystanders as "white-skinned" people and victims as "dark-skinned", we made it explicit that victims and perpetrators belong to different ethnic groups. Similarly, it has been shown that people tend to consider derogatory expressions more inappropriate when stated by a man rather than a woman (Fasoli et al., 2015); therefore, we included female and male perpetrators in the vignette battery. Using cartoons made it easier to take all those considerations into account.

While hate speech is a complex and multifaceted phenomenon, definitions vary across scholarly, legal, psychological and cultural contexts. Legal frameworks, such as the European Union's definition, understand hate speech as all types of expression that incite, promote, spread or justify violence, hatred or discrimination against a person or group of persons or that denigrates them because of their actual or attributed personal characteristics or status such as "race", colour, language, religion, nationality, national or ethnic origin, age, disability, sex, gender identity and sexual orientation (Council of Europe, 2022), emphasising that it incites violence or hatred against a particular group, contributing to a broader "social harm" (Waldron, 2012), and distinguish between speech that expresses offensive ideas and speech that directly incites harm. Moreover, cultural norms also play a significant role in shaping perceptions of hate speech. Activist movements and community standards contribute to an evolving understanding of what constitutes offensive or harmful expression. Finally, from a psychological and sociological perspective, hate speech perpetuates prejudice and discrimination through stereotyping and prejudice manifestation through communication (Fiske, 1998). It inflicts emotional and psychological harm, extending the conversation beyond legalistic definitions (Matsuda, 2018).

Still far from a consensual definition of hate speech incidents (Anderson, Barnes (2022); Lepoutre et al., 2023), in this study, following a legal perspective, we characterise them as those performed by a perpetrator with a degrading and discriminatory intention towards a victim based on a particular personal characteristic (race or ethnic origin, religion, gender, physical or

mental conditions, among others) of the latter (Zapata and Deroy, 2023). As our study focuses on racist hate-speech, the verbal expressions we presented to participants consist of generic, demeaning and discriminatory phrases targeting dark-skinned victims that send a symbolic message that they are unwelcome (Waldron, 2012) and unworthy of social respect (e.g., "You are making our country sick", "Go back home!"). To further select the phrases under assessment, we ran a pilot survey where we tested several common hate expressions and only included as stimuli those that were rated as similarly harmful. With all these measures, we aimed to minimise confounding variables that might otherwise interfere with the research focus of our study.

Experiment 1

Study description. In this first experiment, we investigated the effect of bystanders' silent response when facing a hate speech incident. We collected participants' responses regarding two dependent variables: (1) the incident's perceived level of harm and (2) the blame assigned to the perpetrator.





Regarding the latter, we concretely wondered whether people would consider silent bystanders to contribute to the damage caused by the perpetrator and blame them for their passive response. Following the distribution of responsibility principle (El Zein et al., 2019; Keshmirian et al., 2022), we assumed that if participants blame silent bystanders, they would distribute the responsibility for the harmful outcome between them and the perpetrator and, therefore, assess the perpetrator as less blame-worthy in scenarios with silent bystanders present.

Distributing responsibility refers to dividing decision-making responsibilities among multiple individuals or groups in a collective decision-making process. It is crucial for fostering effective and accountable decision-making in a collaborative setting. Thus, in the context of a perpetrator attacking a victim by using hate speech in front of passive, silent bystanders, this principle allows us to hypothesise that if participants view silent, passive bystanders as perpetrator supporters, they will divide the blameworthiness for taking part in such an act amongst them.

In a within-subjects design, we tested 4 non-factorial experimental conditions. Table 1 lists these conditions, which we refer to as Scenario A, B, C and D. Scenarios A and B are individual scenarios and show incidents that occurred in front of a single bystander. Scenarios C and D are collective scenarios and show incidents that occurred in front of a group of three bystanders. This non-factorial design aimed to compare the effect of an individual remaining silent (Scenario A) to one voicing opposition (Scenario B), but also to test the impact of a bystander staying silent in collective settings, either following the majority reaction (Scenario C) or going against it (Scenario D).

Finally, as an exploratory question, we investigated whether people identify bystanders who witnessed a hate speech incident and who remain silent as implicitly supporting the perpetrator. To this end, we collected participants' responses regarding the number of perpetrator supporters they identified in each scenario.

We formulated the following hypotheses:

Bystanders' reactions	Type of scenario	N° Silent bystanders	N° Opposing bystanders
A 	Individual	1	0
B 	Individual	0	1
C 	Collective	3	0
D 	Collective	1	2

H1: An individual scenario with a silent bystander (Scenario A) will be assessed as more harmful than one with an opposing bystander (Scenario B).

H2: A collective scenario with more silent bystanders present (Scenario C with 3 silent bystanders) will be assessed as more harmful than one with fewer (Scenario D with 1 silent bystander).

H3: In the individual scenario with a silent bystander (Scenario A), the perpetrator will be assessed as less blameworthy than in that with an opposing bystander (Scenario B).

H4: In the collective scenario with more silent bystanders (Scenario C with 3 silent bystanders), the perpetrator will be less blameworthy than those with fewer (Scenario D with 1 silent bystander).

Methods

Participants. We conducted a power calculation with G*Power software for a Friedman test (equivalent to a nonparametric repeated measures ANOVA) and a post hoc Wilcoxon Signed Rank test. Results showed that to detect an estimated small effect size of 0.15 with an alpha probability of 0.05 and a power of 0.80, 62 participants were required for the Friedman and 290 for the Wilcoxon. We then recruited 353 British English-speaking participants through Cloud Research (Amazon Mechanical Turk). We recruited only British participants since the UK is a leader in combating hate speech and creating social awareness about verbal harm. British legal system was a pioneer in implementing hate speech regulations in Western Europe, dating back to the seventeenth century (Rosenfeld, 2003). Therefore, we expected British citizens would be more aware of the effects of showing opposition or remaining silent when facing a hate speech incident (Zapata and Deroy, 2023).

Before the analysis, we excluded data from 24 participants: Six failed the attention check, 17 submitted incomplete surveys, and one submitted a duplicated data set. The final sample size included in the study was $N = 329$ participants (114 female, 210 male, 5 prefer not to say/non-binary).

Procedure. We conducted the study using the Qualtrics online platform (www.qualtrics.com). After providing informed consent, participants were shown four experimental scenarios and one attention check scenario (see below). Participants were asked to rate all scenarios (see Table 1) regarding the incident's level of harm ("In your opinion, how harmful is the situation described above?") and the perpetrator's deserved blame ("To what extent should [perpetrator] A be blamed for the situation described above?"). The order of presentation of these DVs was randomised. Participants assessed all scenarios using a 7-point Likert scale ranging from 0 (Not at all) to 6 (Extremely).

Participants were instructed to respond regarding the (overall) incident level of harm. Additionally, in the instructions we highlighted our interest in participants' personal judgements as ordinary citizens. Our method aligns with many experimental moral philosophy studies (Alfano et al., 2022) focused on capturing participants' general moral judgements. The folk concept of harm captures negative consequences for the individual deployed in moral judgement (Schein, Gray (2018)), which are broader than the concept of pain as specific neural activation (Eisenberger, 2015). Such a concept also includes negative consequences for one's well-being caused by speech, as has been recognised by legal scholars (Petersen, 2016). This is compatible with participants having different understandings of harm. In that sense, we embrace moral pluralism through varieties of values and a flexible conception of "perceived harm" that welcomes diverse perceptions of norm violations and negative affect created among individuals (Schein, Gray (2018)).

Our models are sensitive to participants as a factor due to the inclusion of a per-participant random intercept, which accounts for the fact that some participants may have had higher overall ratings than others.

Additionally, we presented participants with a question regarding the number of perpetrator supporters they identified in each scene ("How many [perpetrator] A's supporters do you identify?"). Participants responded using a forced-choice list that offers "zero", "one", and "two or more" as response options. All visual scenarios and their respective questions were shown in a randomised order. Participants finished the study by answering basic demographic questions. All participants who completed the survey and did not fail the attention check were paid 1.50 USD for a maximum of 8 min of work.

Testing materials (visual vignettes). We created a series of 16 colourful cartoons with a similar structure: All characters appear in a public space (A park, bus stop, street, or subway). A white-skinned perpetrator with an angry face yells a racist remark to a dark-skinned victim (e.g., "Go back home. We do not want your kind here!"), in front of one or three bystanders who witness the incident and either voice their opposition (e.g., "Enough! Stop saying that!") or remain silent. We aimed for consistency in the facial expressions, body language and skin colour of the perpetrators and victims. Perpetrators are angry-faced and show disdain and dislike for the victims; the victims appear alone and look intimidated or ashamed. The bystanders had a direct line of sight to the attack and were close to it. The scenarios were gender-balanced, with female and male perpetrators, victims and bystanders.

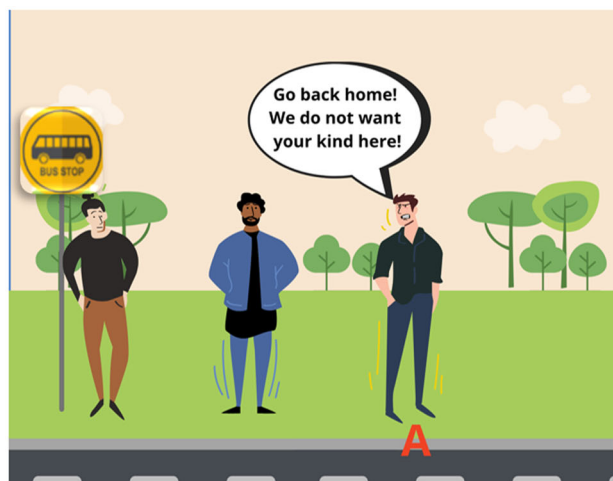
Examples of the visual vignettes used are shown in Fig. 1 (See Supplementary Information section for the complete battery of testing materials).

Attention-check task. An attention check appeared randomly throughout the experiment to ensure participants observed the experimental vignettes and read the questions (Fig. 2). The attention check had a vignette format. Still, it showed a friendly conversation between two people. Participants had to respond by assessing the incident as low in harm and the perpetrator as low in blame (below two on a 7-point Likert scale) to pass the attention check.

Analysis strategy. Data were pre-processed by excluding participants who failed the attention check. As we worked with Likert scales and ordinal data, we conducted a nonparametric Friedman test and a Wilcoxon signed rank test to analyse the differences in participants' median ratings (Sullivan and Artino, 2013) on the two dependent variables (*The incident's level of harm* and *the deserved blame for perpetrators*), across the four experimental conditions. All data analyses were performed in RStudio.

Results

The incident's level of harm. As expected, the results of a non-parametric Friedman test revealed significant differences in the ratings of the incident's level of harm between the experimental conditions ($\chi^2(3) = 27.06$, $p < 0.001$, Kendall's $W = 0.03$ [0.01, 0.05]). However, post hoc testing with Wilcoxon signed rank tests (and Holm-corrected p -values) revealed that there were no significant differences between scenarios A and B (both medians = 6, $r = 0.100$, $p_{adj.} = 0.477$), which rejects our first hypothesis (H1). In individual scenarios where a single bystander witnessed the attack, participants assessed the incident as similarly harmful, independently of whether the bystander showed opposition or remained silent. However, in collective scenarios



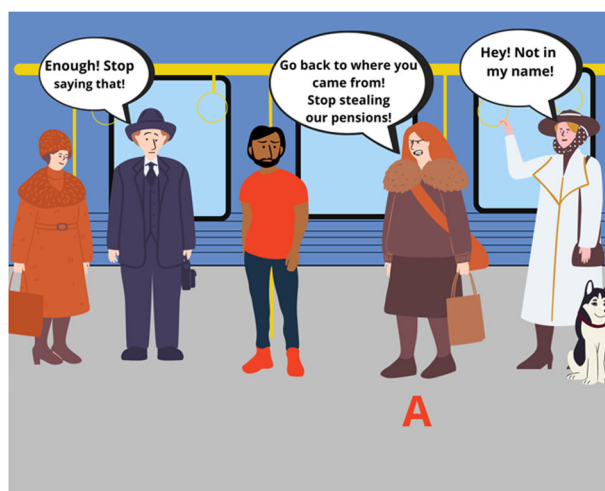
Scenario A



Scenario B



Scenario C



Scenario D

Fig. 1 Visual vignettes (Experiment 1). Example visual vignettes for each of the 4 experimental conditions (Scenarios A-D). The perpetrator is labelled as “A”.



Fig. 2 Attention check. The attention check vignette (Experiment 1).

with three bystanders present, participants assessed the scenario with more silent bystanders (Scenario C) as more harmful than that with fewer (Scenario D), confirming our second hypothesis (H2, Fig. 3a, b).

In addition, we found significant differences between scenario D (median rating = 5) and all other scenarios (all other medians = 6; D vs A $r = 0.262$, $p_{adj.} < 0.001$; D vs B $r = 0.162$, $p_{adj.} = 0.015$; D vs C $r = 0.206$, $p_{adj.} < 0.001$). Thus, with more opposing bystanders present, Scenario D was assessed as the least harmful. Our results show that participants perceived bystander responses as beneficial only in collective settings.

The perpetrators’ deserved blame. Here, the analysis showed that the rating scores for the perpetrator’s blameworthiness were not significantly different among scenarios. The results of a nonparametric Friedman test contradicted hypotheses 3 and 4 and revealed that median ratings for blame were not significantly different ($\chi^2(3) = 5.71$, $p = 0.127$, Kendall’s $W = 0.006$ [0.001, 0.02]). Participants blamed perpetrators similarly, disregarding whether they attacked the victim in front of silent or opposing bystanders and whether the attack occurred in individual or collective settings (Fig. 3c, d).

Exploratory analysis. We explored whether people tend to identify silent bystanders as supporting the perpetrator. To do so, we conducted a Cumulative Link mixed model regression analysis to test whether the number of bystanders present predicts the number of perpetrator supporters identified. We found that the

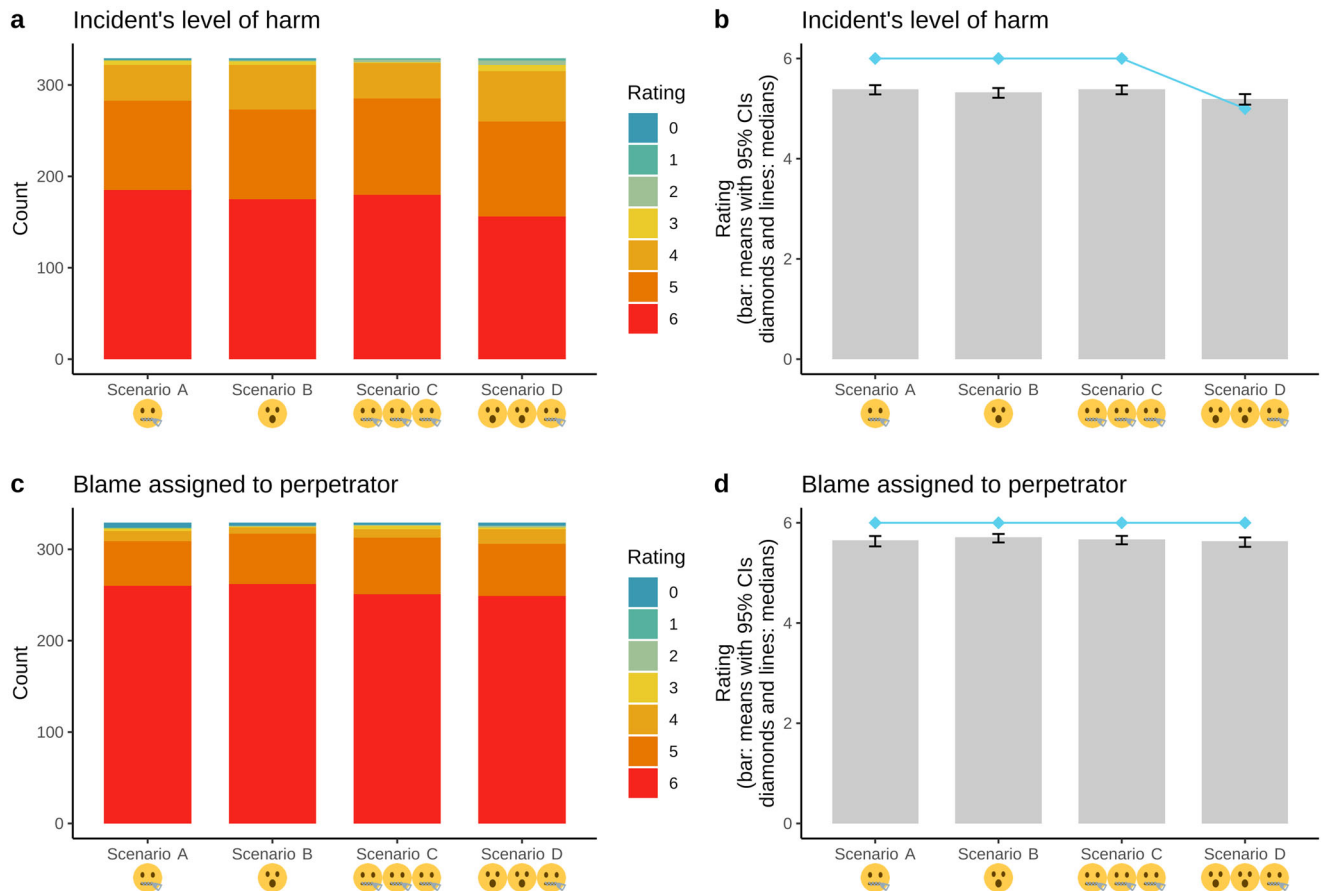


Fig. 3 Results (Experiment 1). **a** A stacked bar chart showing the distribution of ratings for the incident's level of harm, grouped by scenario; **(b)** grey bars show mean rating (and whiskers show 95% bootstrapped CIs) for the incident's level of harm; blue diamonds and lines show median responses; **(c)** A stacked bar chart showing the distribution of ratings for the blame assigned to perpetrators, grouped by scenario; and **(d)** grey bars show mean rating (and whiskers show 95% bootstrapped CIs) for the blame assigned to perpetrators, blue diamonds and lines show median responses.

number of silent bystanders was a significant positive predictor of perpetrator supporters identified ($b = 0.60 [0.43, 0.77], SE = 0.08, t = 7.02, p < 0.001$). Scenario C, with three silent bystanders present, was rated as having the highest number of perpetrator supporters (Fig. 4).

However, our design did not address whether—when counting perpetrator supporters—participants considered only silent bystanders or considered the silent victim too. Therefore, we address the issue of silent vs opposing responses in a more controlled manner in Study 2.

Discussion on Experiment 1

Experiment 1 showed that bystanders' reactions affected the perception of the harm caused by a hate speech attack only in collective settings when other bystanders are shown. This might suggest that when we do not offer participants enough elements to intuit the social norm against hate speech (by showing them how other bystanders react), they evaluate both hate incidents as similarly harmful, independently of whether the bystander present responded by remaining silent or showing opposition. Additionally, our results suggested that people evaluate scenarios where a group of bystanders voiced their opposition as less harmful than those where a group remained silent (Fig. 3a, b).

However, it remains unclear under precisely which conditions the perception of harm was affected by bystanders' opposing a hate speech attack in collective settings: Does opposing hate speech against the social norm—when the majority remains silent—affect the

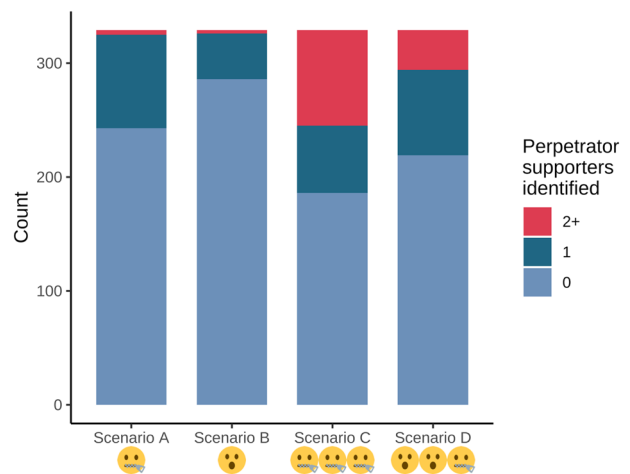






Fig. 4 Results of the exploratory analysis. Graph showing the perpetrator supporters identified grouped by scenarios.

perceived damage differently than opposing it following the majority? We ran a second experiment to answer these questions.

Experiment 2

Study description. Based on Experiment 1's findings, we changed to focus exclusively on collective settings, where there

Table 2 Experimental conditions tested in Experiment 2, with target bystanders indicated with arrows.

Experimental conditions		Majority response	
		Remain silent	Show opposition
Target-bystander response	Remain silent	A → 	C → 
	Show opposition	B → 	D → 

were always three bystanders. We varied the number of opposing responses from zero to three (of three total bystanders). This meant that opposition could be absent (0/3), be a minority response (1/3), be a majority response (2/3), or be unanimous (3/3). Additionally, we designated one of the bystanders the “target bystander” so that the questions could focus participants’ attention on a specific bystander, and we could thus ask participants to rate the specific target’s contribution to overall harm during the hate speech incident. The target bystander could be silent or opposing, with or against the majority. Accordingly, this yields a factorial design combining two independent variables (IVs). IV1 (“target response”) the target-bystander’s response to the hate speech incident with two levels: showing opposition vs remaining silent; and IV2 (“majority response”) the response of the bystanders majority also with two levels: showing opposition vs remaining silent. Table 2 lists all the conditions. Target bystanders, which are the focus of questions about specific bystanders’ contributions to harm, are indicated with an arrow (→).

The factorial design allowed us to test individual (target bystander) and collective (group of bystanders) reactions to hate speech incidents and simultaneously to test the effect of the target bystander responding with or against the majority. As shown in Table 2, the target bystander remained silent in scenarios A and C. However, in scenario A, she did it jointly with all other bystanders, while in C, she remained silent when the majority showed opposition to the hate speech incident. Likewise, in scenarios B and D, the target bystander opposed the attack. Still, in scenario B, she opposed the attack when the majority remained silent. In contrast, in scenario D, she opposed the hate attack together with the rest of the bystanders.

A within-subjects design allowed all participants to evaluate four experimental conditions with zero, one, two or three opposing bystanders (referred to as scenarios A, B, C and D). We collected two dependent variables: the overall level of harm of the incident (“harm” DV1) and the specific contribution of the target bystander to that harm (“contribution” DV2). For the latter, we asked participants whether the target bystander’s response increased or decreased the harm caused by the incident. The order of presentation of these questions was randomised.

We tested the following hypotheses about how bystander responses will affect the perception of the harm caused by a hate speech incident:

H1: *Target-bystander’s opposing response* will contribute negatively to (i.e., reduce) the perceived harm. (DV2 as a function of IV1)

H2: When most bystanders remain silent, a *target bystander opposing the attack* will reduce the perceived harm less than when the others oppose the attack (DV2 as a function of IV1 × IV2).

H3: *The level of harm of the incident* will be reduced accordingly to the number of opposing bystanders present. (DV1 as a function of the number of bystanders)

H4: *The level of harm of the incident* will be reduced when showing opposition is the majoritarian reaction among bystanders (DV1 as a function of IV2, indicating a social norm).

H5: *The level of harm of the incident* will be reduced when showing opposition is unanimous among bystanders (DV1 as a function of unanimity, indicating a robust social norm).

Methods

Participants. As we planned to analyse the DVs using cumulative link mixed-effects models, we conducted a power calculation through simulation for mixed models with the mixed-power R package (Kumle, Vö and Draschkow, 2018). In addition, we used pilot data to obtain estimates for fixed and random effects. Results showed that to reach a power of 0.80, 225 participants were required.

We recruited 272 British English-speaking participants through Prolific (www.prolific.co). Prior to analysis, data from four participants who failed the attention checks (see below) were removed. The final sample size included in the study was $N = 269$ participants (134 female, 2 prefer not to say/non-binary).

Testing materials. As for Experiment 1, we created colourful cartoons representing hate speech incidents. However, this time all four scenarios were collective (group) scenarios of three bystanders showing opposition or remaining silent. Examples of the visual vignettes are shown in Fig. 5. In addition, in each scenario, there was a “target” bystander who either appeared silent or showed opposition, with or against the majority (See Table 2). Figure 6 shows, as an example, the target bystander in Scenario D, who appears to show opposition in line with the majority. We showed participants a vignette showing only the target bystander when we asked them to assess a specific target’s contribution to overall harm during the hate speech incident (See Supplementary Information section for the full battery of visual vignettes).

Attention-check task. As in Experiment 1, we used an attention check that appeared randomly in the trial order. It showed 3 characters, two of whom were talking friendly. We asked participants how many people were speaking in the scene, and they had to respond 2 on a 7-point Likert scale to pass the attention check.

Procedure. We conducted the study using the Qualtrics software (www.qualtrics.com). After giving consent, participants were shown four experimental scenarios and one attention check in random order. In each of the experimental trials, a perpetrator shouts a hateful remark towards a victim in the presence of a group of three bystanders who respond individually or collectively, each remaining silent or voicing their opposition against the attack, as shown in Table 2.

Participants were asked to rate all scenarios regarding the incident’s overall level of harm (“*In your opinion, how harmful is the incident shown above?*”, DV1). As in Experiment 1, responses were on a 7-point Likert scale ranging from 0 (Not at all) to 6 (Extremely). In addition, we asked them to rate a target-bystander’s individual contribution to the harm caused by the incident (“*To what extent does this person’s reaction contribute to the harm caused by the incident. If you consider his reaction plays no role, please, place the cursor on zero.*”, DV2). To answer this question, we presented participants with a picture of a target bystander (Fig. 6), and they responded using a bipolar 7-point Likert scale ranging from -3 (Reduces the harm) to 3 (Increases the harm). The middle point (0) was explicitly labelled as



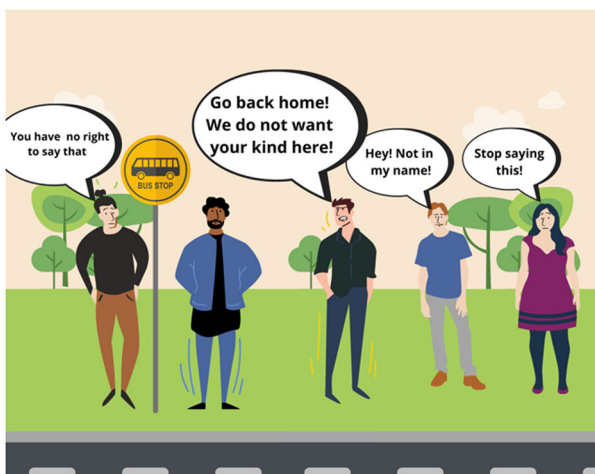
Scenario A



Scenario B



Scenario C



Scenario D

Fig. 5 Visual Vignettes (Experiment 2). The image shows example visual vignettes for each of the 4 experimental conditions: Scenario A with 0 opposers, Scenario B with one, Scenario C with 2 and Scenario D with 3 opposers.

“neutral” to highlight to participants that this means “had no effect on overall harm”.

Participants finished the study by answering basic demographic questions. All participants who completed the survey and did not fail the attention check were paid £0.75 for a maximum of 5 min of work.

Analysis strategy. First, we pre-processed the data by excluding participants who failed the attention check. Then, we ran a series of cumulative link mixed-effects regressions (CLMM, R package “ordinal”, Christensen, 2022) to test the hypotheses. All data analyses were performed in R. The OSF repository for this study (https://osf.io/nfyg9/?view_only=3fe4e0bf7ddd41d4a27dc252cfb67455) contains the data and analyses. The models reported below (Table 3) were not pre-registered, but the full R analysis script is available in the above study repository. Regression coefficients are reported with 95% confidence intervals (CIs).

Results

The specific contribution of the target bystander to that harm. First, we tested the effect of a target-bystander reaction (showing opposition or remaining silent) on perceived harm.

To do so, we ran a cumulative link mixed-effects regression (Model 1) with the target bystander’s contribution to the harm caused (reduce or increase) as the outcome variable and the target-bystander reaction (opposing or remaining silent) as the predictor. Results showed that the target-bystander response significantly and negatively predicted harm (i.e., reduced it) when the target bystander opposed the attack ($b = -3.57 [-3.90, -3.24]$, $SE = 0.17$, $t = -21.03$, $p < 0.001$), confirming H1.

We ran a cumulative link mixed-effects regression (Model 2) with the target bystander’s contribution to the harm caused (reduce or increase) as the outcome variable, with this regressed on the target-bystander reaction (remain silent or show opposition), the social norm followed by the majority of bystanders (opposing or remaining silent), and an interaction term. Results showed a nonsignificant effect of the social norm (remaining silent: $b = -0.01 [-0.32, 0.29]$, $SE = 0.16$, $t = -0.07$, $p = 0.947$); a significant negative effect of showing opposition as target-bystander reaction (reducing harm: $b = -3.23 [-3.61, -2.84]$, $SE = 0.19$, $t = -16.34$, $p < 0.001$), and a significant interaction between showing opposition as the targeted-bystander reaction and remaining silent as the social norm: ($b = -0.83 [-1.26, -0.39]$, $SE = 0.22$, $t = -3.68$, $p < 0.001$). Thus, the target bystander’s opposition to the attack

reduces harm more when it goes against a social norm of being silent, counter to H2, which predicted the opposite effect.

Using the R package “performance”, we tested the fit of both previous regression models as indexed by the Bayesian Information Criterion (BIC). The results indicated that Model 2 fits the data better than Model 1 (BIC model 1 = 3236, BIC model 2 = 3223, ΔBIC = 13, weight favouring model 2 = 0.9989). Thus, the best available description of the data is that participants perceived the harm-reducing effect as higher in Scenario B, where a single bystander shows opposition while all the others remain silent (Fig. 7a, b).

The overall level of harm of the incident. Secondly, we tested—again, always in collective settings—the effect of several predictors (number of opposing bystanders {0, 1, 2 or 3}, a majority opposition response and a unanimous opposition response) on participants’ perceptions of the overall harm caused to victims. For this purpose, we ran three different cumulative link mixed model regressions.

Model 3 regressed the incident’s overall perceived harm on the number of opposing bystanders and showed a significant negative

effect of the number of opposers ($b = -0.16 [-0.157, -0.156]$, $SE < 0.001$, $t = -613.42$, $p < 0.001$). Model 4 had the same outcome variable but regressed this on the majority bystander response (social norm) and showed a nonsignificant effect when the majority opposed ($b = -0.24 [-0.53, 0.05]$, $SE = 0.15$, $t = -1.64$, $p = 0.102$). Finally, Model 5 regressed the same outcome variable on the dichotomous unanimity variable (whether all bystanders opposed or not, with the former reflecting a robust social norm). The results showed a significant negative effect when all bystanders opposed ($b = -0.63 [-0.97, -0.30]$, $SE = 0.17$, $t = -3.74$, $p < 0.001$).

Lastly, using the “performance” package, we evaluated the fit of the three previous regression models (Model 3 BIC = 2077, model 4 BIC = 2071, model 5 BIC = 2067, ΔBIC = 4 for model 5 vs next-best model 4, weight in favour of model 5 = 0.831). Thus, the best available description of the data is that the incident’s overall level of harm is better reduced when the opposition against a hate speech incident is unanimous among bystanders, thereby becoming a robust social norm (Fig. 7c, d).

Discussion on Experiment 2

Experiment 2 placed a given bystander’s response to a hate speech incident in the context of other bystanders’ reactions (reflecting overall levels of opposition/social norms). Results show that participants, as third-party observers, judged that remaining silent could increase the perceived harm of a hate speech incident, that a given individual’s speaking out is more impactful when the majority of bystanders are silent. Crucially, however, the best way to reduce harm overall is to have a robust social norm (followed unanimously) in favour of speaking out against hate speech. Thus, assessing a bystander’s response to hate speech without considering the social context (and any empirical social norms in place) could overestimate its impact on perceived harm. As Fig. 7 shows, the variation in the incident’s overall level of harm is relatively small (Fig. 7d) compared to the variation in how a bystander’s response impacts overall harm (Fig. 7a, b) when it is assessed individually. Moreover, although participants praise single opposers who raise their voices amid the silent majority, our results show that only unanimous opposition significantly reduces the public perception of the harm caused.

General discussion

Experts from different disciplines have strongly advocated for counterspeech as a tool against hate speech and its harmful consequences for victims and society (for an overview, see Cepollaro et al., 2023). In this paper, our starting point was to explore whether those who might counter or block hate speech find voicing opposition helpful in reducing the harm created.

Our results show that ordinary people overlook the effect of a silent or an opposing response in the harm created by hate speech when they assess those reactions as individual responses from a single bystander. Moreover, opposing a hate attack when all other



Fig. 6 Target bystander. The image illustrates the target bystander in Scenario D. Such an image was presented alongside all questions about a bystander’s individual contribution to the overall harm caused by the incident to ensure that participants knew which bystander was the focus of each question.

Table 3 Cumulative link mixed models tested in Experiment 2.

Model	Outcome variable	Predictor variable
1	Target-bystander’s contribution to the harm caused (increases or reduces)	Target-bystander’s reaction (show opposition or remain silent)
2	Target-bystander’s contribution to the harm caused (increases or reduces)	Target-bystander’s reaction (show opposition or remain silent) * social norm (reaction followed by the majority of bystanders)
3	Level of harm of the incident	Number of opposing bystanders (3, 2, 1, 0)
4	Level of harm of the incident	Showing opposition as majority response (Social norm supported by the majority)
5	Level of harm of the incident	Showing opposition as unanimous response (Robust social norm unanimously supported)

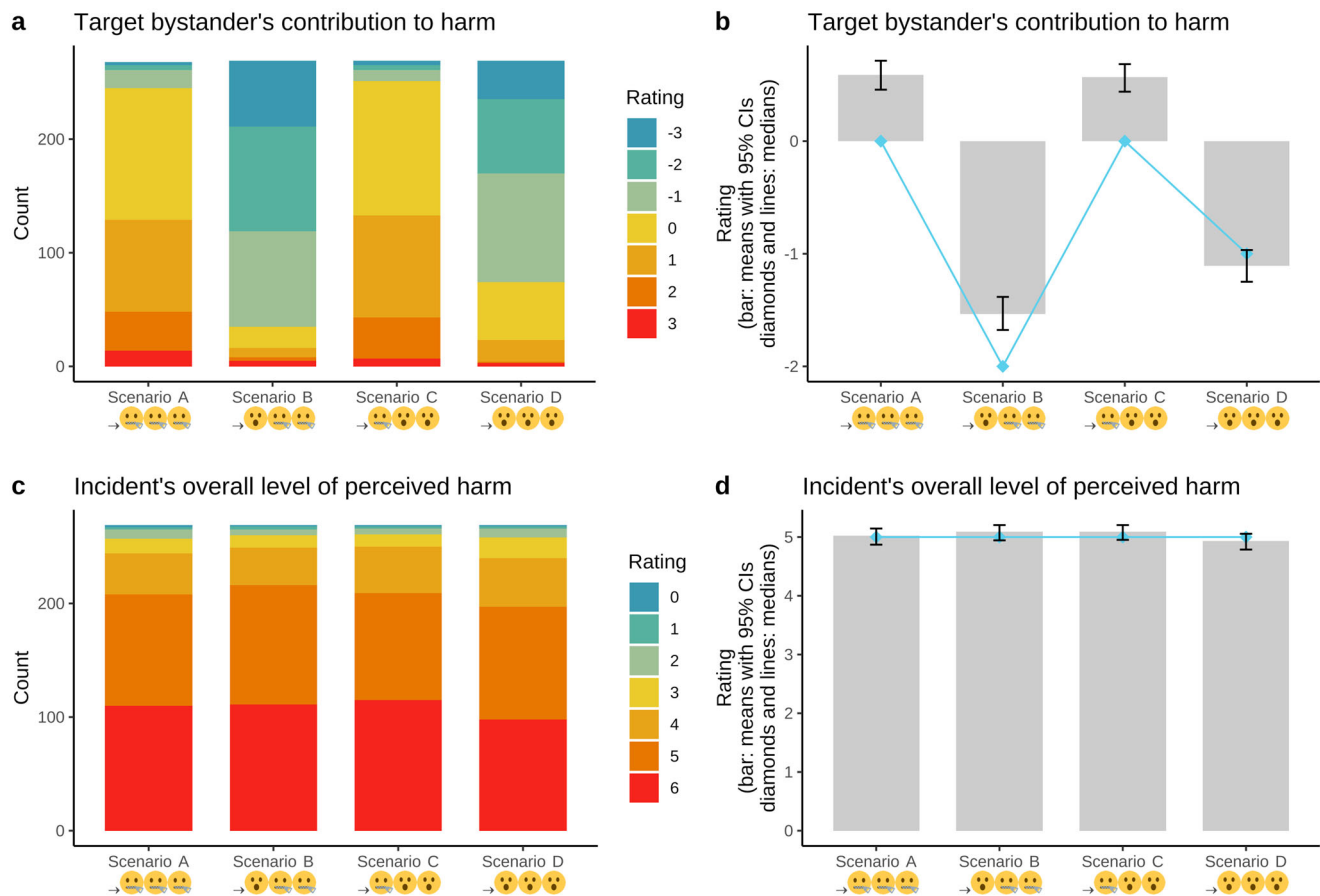


Fig. 7 Results (Experiment 2). Responses are grouped by scenario and the target bystander in each scenario is indicated with an arrow. **a** Stacked bar chart showing the rating distribution for the target bystander's contribution to harm (positive ratings = increase harm, negative ratings = reduce harm, zero = makes no difference). **b** Grey bars show the mean rating (and whiskers show 95% bootstrapped CIs) of the target bystander's contribution (increase or reduce) to the damage caused by the incident; blue diamonds and lines show median responses. **c** Stacked bar chart showing the rating distribution for the incident's overall level of perceived harm. **d** Grey bars show the mean rating (and whiskers show 95% bootstrapped CIs) of the incident's overall level of perceived harm; blue diamonds and lines show median responses.

bystanders keep quiet is seen as more helpful in reducing harm. However, when we offer participants scenes with a social context and a clear social norm against hate speech (followed by most bystanders), they judge that an isolated opposing response does not reduce the perceived harm, though a unanimous collective opposition can do so. Our results support that group responses to hate speech can modulate its damage by indicating either a condoning or a condemning social norm.

Chater and Loewenstein (2022) pointed out that discrimination is a type of social problem, as inequality or misinformation are, in which the phrase “small changes can make a big difference” does not apply. Our results point in the same direction, suggesting that showing opposition against hate speech is ineffective in isolation and that groups need to respond against demeaning and discriminatory speech as a social norm to effectively reduce its harm.

Limitations to generality. As hate speech is highly context-dependent, we conducted our study with only British English-speaking participants; further research is needed to explore whether our findings are replicated with non-English-speaking participants from different countries. Likewise, we only tested racist hate speech with case vignettes representing “real-life” attacks. However, future research can extend our findings by investigating people's responses to hate speech based on different biases (homophobia, transphobia, based on religious hatred, among others) in various settings like online forums. Moreover, it

can investigate correlations between actual victims' perceptions of harm reduction and bystanders' interventions against hate speech.

In addition, our visual stimuli only used counterspeech that confronts perpetrators (e.g., “Stop saying that”, “You have no right to say that”), and further research should explore whether people's responses change if we direct the counter-speech to the victim (e.g., “Don't believe him”, “I welcome you to this country”) or modulate it, making it more indirect (e.g., “I am calling the police”).

Finally, following our account, in forthcoming work, we will test whether using expressions that imply a collective response would reduce the harm better than those that suggest individual responses (e.g., “We welcome you”, “We will report this to the police”, “We don't share that opinion”).

Conclusions

Our study contributes to the literature on hate speech and its perceived harms, beginning to explore the role of counter speech. We examined how ordinary people view an individual bystander's response to hate speech relative to a group response in several social contexts: with no information about the social norm on how to respond to a hate speech incident (i.e., the assessed scene shows an incident with a single bystander present), with the standard being to show opposition (i.e., the scene showed a majority of bystanders opposing the speech attack) and with the

norm being to remain silent (i.e., the scene showed a majority remaining silent when facing a hate speech incident).

In two experiments, we used specially designed visual stories to demonstrate that when people have information about the social norm regarding hate speech, they better understand the impact of intervening as a third-party. This influence is often missed without this context. We showed participants scenes where most bystanders either spoke against hate speech or stayed quiet. Those who saw these scenes considered that individually opposing hate speech can lessen its harm, while not reacting might worsen it. However, when participants lacked the broader social context, they viewed both silent and vocal responses to hate speech as equally effective, failing to recognize the importance of bystander intervention in these situations.

Additionally, our research revealed that the way bystanders react to hate speech significantly influences how others perceive its impact. When most bystanders visibly oppose hate speech, third-party observers tend to see these incidents as less harmful to both individuals and society. We also found that the number of bystanders who either protest or stay silent plays a crucial role in shaping these perceptions. It's not sufficient for just one person to speak out against hate speech; effective responses need to be collective, demonstrating a broader social stance against such behaviour.

Hate speech is better addressed by group responses than individual efforts, and social norms of speaking out against spreading hate require strong (even unanimous) support to modulate the perceived harm. As hate speech is ultimately about demeaning social groups more than specific individuals (Perry and Alvi, 2012), it also requires collective responses reaffirming coexistence within democratic principles of tolerance and respect for diversity. Our findings show that people's intuitions point in the same direction, supporting public policies that promote civic engagement against hate speech. These results are in line with empirical studies showing that signalling condemnation of racism leads people to hold stronger anti-racist opinions whereas hearing others condoning racism causes the contrary effect (Blanchard et al., 1994), even after terrorist attacks (Álvarez-Benjumea and Winter, 2020).

The implications of our findings in moral philosophy are clear: As social beings, when facing incidents that we clearly identify as harmful to others, our responses are not only informed by individual moral principles but also by social ones: an essential source of information on how to respond against hate speech incidents is how others do (Tunçgenç et al., 2021).

Our findings support the philosophical theories defending the importance of showing opposition against hate speech in avoiding the perpetuation of oppressive norms and hierarchies based on xenophobic, racist or similar motives (Ayala and Vasilyeva, 2016; Langton 2018a; 2018b; Lepoutre, 2021; Caponetto and Cepollaro, 2023; Howard, 2021), with an addition concerning the collective character of such a response. But they go further, responding to the existent demands of contextualization and specificity in counter-speech strategies (Cepollaro et al., 2023; Howard, 2021), suggesting a group response appears to third-party observers as more effective in reducing hate speech harms than individual efforts, despite its salience when a majority remain silent.

Moreover, we argue that the damage created by silent bystanders does not come from their being immediately perceived as perpetrator supporters but from increasing the uncertainty about how we, as a society, treat minorities and disfavoured groups. Sometimes, our silence is eloquent in showing opposition (e.g., when we avoid saluting sexist expressions). Still, our responses to harmful practices are highly contextual and, therefore, in times of change and ambivalence, making explicit our

rejection of the introduction of discourses that demean, oppress, and silence others based on who they can make the most difference.

Our findings also highlight the importance of policies that encourage public engagement against hate speech. Social norms that perpetuate silence in the face of social hierarchies and inequality are pervasive but not immutable. They can be challenged and changed (as noted by Bicchieri and de Silva and Simpson). By collectively speaking out against hate speech, groups can challenge the broader norm of silence in the face of identity-based abuse and discrimination, demonstrating that these norms are not universal and that harmonious coexistence in diverse societies is possible.

Data availability

The data sets generated and analysed for this study are available through the Open Science Framework, which also contains the analysis scripts and study stimuli: (https://osf.io/nfyg9/?view_only=3fe4e0bf7ddd41d4a27dc252cfb67455).

Received: 24 July 2023; Accepted: 25 January 2024;
Published online: 29 February 2024

References

- Alfano M, Machery E, Plakias A, Loeb D (2022) Experimental Moral Philosophy. In: Zalta EN and Nodelman U (eds) Stanford Encyclopedia of Philosophy (Fall 2022) Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/fall2022/entries/experimental-moral/>. Accessed 10 Nov 2023
- Almagro M, Hannikainen IR, Villanueva N (2022) Whose words hurt? Contextual determinants of offensive speech. *Personal Soc Psychol Bull* 48(6):937–953. <https://doi.org/10.1177/01461672211026128>
- Álvarez-Benjumea A, Winter F (2020) The breakdown of antiracist norms: a natural experiment on hate speech after terrorist attacks. *Proc Natl Acad Sci* 117(37):22800–22804. <https://doi.org/10.1073/pnas.2007971117>
- Álvarez-Benjumea A (2023) Uncovering hidden opinions: social norms and the expression of xenophobic attitudes. *Eur Sociol Rev* 39(3):449–463. <https://doi.org/10.1093/esr/jcac056>
- Anderson L, Barnes MR (2022) Hate Speech. In: Zalta EN (ed) Stanford encyclopedia of philosophy. Springer <https://plato.stanford.edu/entries/hate-speech/>. Accessed 10 Nov 2023
- Ashburn-Nardo L, Blanchard JC, Petersson J, Morris KA, Goodwin SA (2014) Do you say something when it's your boss? The role of perpetrator power in prejudice confrontation. *J Soc Issues* 70(4):615–636. <https://doi.org/10.1111/josi.12082>
- Ashburn-Nardo L, Lindsey A, Morris KA, Goodwin SA (2020) Who is responsible for confronting prejudice? The role of perceived and conferred authority. *J Bus Psychol* 35(6):799–811. <https://doi.org/10.1007/s10869-019-09651-w>
- Atzmüller C, Steiner PM (2010) Experimental vignette studies in survey research. *Methodology* 6(3):128–138. <https://doi.org/10.1027/1614-2241/a000014>
- Ayala S, Vasilyeva N (2016) Responsibility for silence. *J Soc Philos* 47(3):256–272. <https://doi.org/10.1111/josp.12151>
- Barr A, Lane T, Nosenzo D (2018) On the social inappropriateness of discrimination. *J Public Econ* 164:153–164. <https://doi.org/10.1016/j.jpubeco.2018.06.004>
- Bicchieri C (2005) *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press
- Bicchieri C, Mercier H (2014) Norms and beliefs: how change occurs. In: Xenitidou M, Edmonds B (eds) *The complexity of social norms*. Springer International Publishing, pp 37–54 <https://doi.org/10.1007/978-3-319-05308-0>
- Bicchieri C (2016) *Norms in the wild. How to Diagnose, Measure, and Change Social Norms*. Oxford University Press
- Bicchieri C, Dimant E (2022) Nudging with care: The risks and benefits of social information. *Public choice* 191(3–4):443–464. <https://doi.org/10.1007/s11127-019-00684-6>
- Blanchard FA, Crandall CS, Brigham JC, Vaughn LA (1994) Condemning and condoning racism: a social context approach to interracial settings. *J Appl Psychol* 79(6):993–997. <https://doi.org/10.1037/0021-9010.79.6.993>
- Boushehrian M (2020) *Hate crime in the UK: comparing hate crime perception in ethnic groups with regards to their socio-economic status*. Dissertation, University of Portsmouth. <https://doi.org/10.13140/RG.2.2.31296.05127>

- Caponetto L, Cepollaro B (2023) Bending as counterspeech. *Ethical Theory Moral Pract* 26(4):577–593. <https://doi.org/10.1007/s10677-022-10334-4>
- Cepollaro B, Lepoutre M, Simpson RM (2023) Counterspeech. *Philosophy Compass*, 18(1). <https://doi.org/10.1111/phc3.12890>
- Chater N, Loewenstein G (2022) The i-frame and the s-frame: how focusing on individual-level solutions has led behavioral public policy astray. *Behav Brain Sci* 1–60. <https://doi.org/10.1017/S0140525X22002023>
- Christensen R (2022) ordinal—Regression Models for Ordinal Data (R package version 2022.11-16). <https://cran.r-project.org/package=ordinal>. Accessed 10 Nov 2023
- Cook WL, Sheppard L (2018) Not doing nothing: third parties' cognitive reactions to mistreatment of others. *Acad Manag Proc* 2018(1):15718. <https://doi.org/10.5465/AMBPP.2018.15718abstract>
- Council of Europe (2022) Recommendation CM/Rec(2022)16 of the Committee of Ministers to member States on combating hate speech. https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=0900001680a67955#_ftn2. Accessed 10 Nov 2023
- Czopp AM, Monteith MJ (2003) Confronting prejudice (literally): reactions to confrontations of racial and gender bias. *Personal Soc Psychol Bull* 29(4):532–544. <https://doi.org/10.1177/0146167202250923>
- de Araujo E, Altay S, Bor A, Mercier H (2022) Dominant jerks: people infer dominance from the utterance of challenging and offensive statements. *Soc Psychol Bull* 16(4). <https://doi.org/10.32872/spb.6999>
- de Silva A, Simpson RM (2022) Law as counterspeech. *Ethical Theory Moral Pract*. <https://doi.org/10.1007/s10677-022-10335-3>
- Dessel AB, Goodman KD, Woodford MR (2017) LGBT discrimination on campus and heterosexual bystanders: understanding intentions to intervene. *J Divers High Educ* 10(2):101–116. <https://doi.org/10.1037/dhe0000015>
- Dickter CL, Newton VA (2013) To confront or not to confront: non-targets' evaluations of and responses to racist comments. *J Appl Soc Psychol* 43:E262–E275. <https://doi.org/10.1111/jasp.12022>
- Diekmann A, Przepiorka W, Rauhut H (2015) Lifting the veil of ignorance: An experiment on the contagiousness of norm violations. *Rationality and Society* 27(3):309–333. <https://doi.org/10.1177/1043463115593109>
- Eisenberger NI (2015) Social pain and the brain: controversies, questions, and where to go from here. *Annu Rev Psychol* 66(1):601–629. <https://doi.org/10.1146/annurev-psych-010213-115146>
- El Zein M, Bahrami B, Hertwig R (2019) Shared responsibility in collective decisions. *Nat Hum Behav* 3(6):554–559. <https://doi.org/10.1038/s41562-019-0596-4>
- Fasoli F, Carnaghi A, Paladino MP (2015) Social acceptability of sexist derogatory and sexist objectifying slurs across contexts. *Lang Sci* 52:98–107. <https://doi.org/10.1016/j.langsci.2015.03.003>
- Fiske ST (1998) Stereotyping, prejudice, and discrimination. In: Gilbert DT, Fiske ST, Lindzey G (eds) *The handbook of social psychology*. McGraw-Hill, pp 357–411
- Fumagalli C (2021) Counterspeech and ordinary citizens: how? when? *Polit Theory* 49(6):1021–1047. <https://doi.org/10.1177/0090591720984724>
- Gagliardone I, Gal D, Alves T, Martinez G (2015) Countering online hate speech. Unesco Publishing. <https://unesdoc.unesco.org/ark:/48223/pf0000233231>. Accessed 10 Nov 2023
- Gelber K (2012) Reconceptualizing counterspeech in hate speech policy (with a Focus on Australia). In *The Content and Context of Hate Speech*. Cambridge University Press, pp 198–216. <https://doi.org/10.1017/CBO9781139042871.016>
- Gelber K, McNamara L (2016) Evidencing the harms of hate speech. *Soc Identities* 22(3):324–341. <https://doi.org/10.1080/13504630.2015.1128810>
- Gibson JL, Epstein L, Magarian GP (2020) Taming uncivil discourse. *Polit Psychol* 41(2):383–401. <https://doi.org/10.1111/pops.12626>
- Goldberg SC (2010) The epistemology of silence. In: *Social Epistemology*. Oxford University Press, pp 243–261. <https://doi.org/10.1093/acprof:oso/9780199577477.003.0012>
- Goldberg SC (2020) *Conversational Pressure*. Oxford University Press
- Greenberg J, Pyszczynski T (1985) The effect of an overheard ethnic slur on evaluations of the target: How to spread a social disease. *J Exp Soc Psychol* 21(1):61–72. [https://doi.org/10.1016/0022-1031\(85\)90006-X](https://doi.org/10.1016/0022-1031(85)90006-X)
- Gulker JE, Mark AY, Monteith MJ (2013) Confronting prejudice: the who, what, and why of confrontation effectiveness. *Soc Influ* 8(4):280–293. <https://doi.org/10.1080/15534510.2012.736879>
- Henry PJ, Butler SE, Brandt MJ (2014) The influence of target group status on the perception of the offensiveness of group-based slurs. *J Exp Soc Psychol* 53:185–192. <https://doi.org/10.1016/j.jesp.2014.03.012>
- Holm G, Sahlström F, Zilliacus H (2018) Arts-based visual research. In: Leavy P (ed) *Handbook of arts-based research*. Guilford Press, pp 311–335
- Hornsey MJ, Imani A (2004) Criticizing groups from the inside and the outside: an identity perspective on the intergroup sensitivity effect. *Personal Soc Psychol Bull* 30(3):365–383. <https://doi.org/10.1177/0146167203261295>
- House BR (2018) How do social norms influence prosocial development? *Curr Opin Psychol* 20:87–91. <https://doi.org/10.1016/j.copsyc.2017.08.011>
- Howard JW (2021) Terror, hate and the demands of counter-speech. *Br J Political Sci* 51(3):924–939. <https://doi.org/10.1017/S000712341900053X>
- Janson GR, Hazler RJ (2004) Trauma reactions of bystanders and victims to repetitive abuse experiences. *Violence Vict* 19(2):239–255. <https://doi.org/10.1891/vivi.19.2.239.64102>
- Janson GR, Carney JV, Hazler RJ, Oh I (2009) Bystanders' reactions to witnessing repetitive abuse experiences. *J Couns Dev* 87(3):319–326. <https://doi.org/10.1002/j.1556-6678.2009.tb00113.x>
- Keshmirian A, Hemmatian B, Bahrami B, Deroy O, Cushman F (2022) Diffusion of punishment in collective norm violations. *Sci Rep*. 12(1):15318. <https://doi.org/10.1038/s41598-022-19156-x>
- Khanolainen D, Semenova E (2020) School bullying through graphic vignettes: developing a new arts-based method to study a sensitive topic. *Int J Qual Methods*, 19. <https://doi.org/10.1177/1609406920922765>
- Kumle L, Vö MLH, Draschkow D (2018) Mixedpower: a library for estimating simulation-based power for mixed models in R (1.0). Zenodo. <https://doi.org/10.5281/zenodo.1341048>
- Langton R (2007) Disenfranchised silence. In: Brennan G, Goodin R, Jackson F, Smith M (eds) *Common minds: themes from the philosophy of philip pettit*. Oxford University Press, pp 199–214
- Langton R (2012) Beyond belief: pragmatics in hate speech and pornography. In: McGowan MK and Maitra I (eds) *Speech and Harm: Controversies Over Free Speech*. Oxford University Press, pp 144–164. <https://doi.org/10.1093/acprof:oso/9780199236282.003.0004>
- Langton R (2018a) The authority of hate speech. In: Gardner J, Green L, Leiter B (eds) *Oxford Studies in Philosophy of Law*, vol 3. Oxford University Press, pp 132–152. <https://doi.org/10.1093/oso/9780198828174.003.0004>
- Langton R (2018b) Blocking as counter-speech. In: Fogal D, Harris DW, Moss M (eds) *New Work on Speech Acts*. Oxford University Press, pp 1–36. <https://doi.org/10.1093/oso/9780198738831.003.0006>
- Leets L (2001) Explaining perceptions of racist speech. *Commun Res* 28(5):676–706. <https://doi.org/10.1177/009365001028005005>
- Leonhard L, Ruef C, Obermaier M, Reinemann C (2018) Perceiving threat and feeling responsible. How severity of hate speech, number of bystanders, and prior reactions of others affect bystanders' intention to counterargue against hate speech on Facebook. *Stud Commun Media* 7(4):555–579. <https://doi.org/10.5771/2192-4007-2018-4-555>
- Lepoutre M (2017) Hate speech in public discourse. *Soc Theory Pract* 43(4):851–883. <https://doi.org/10.5840/soctheorpract201711125>
- Lepoutre M (2019) Hate speech laws: expressive power is not the answer. *Leg Theory* 25(4):272–296. <https://doi.org/10.1017/S135232522000004X>
- Lepoutre M (2021) *Democratic speech in divided times*. OUP: Oxford University Press
- Lepoutre M, Vilar-Lluch S, Borg E, Hansen N (2023) What is hate speech? The case for a corpus approach. *Crim Law Philos*. <https://doi.org/10.1007/s11572-023-09675-7>
- Maitra I (2004) Silence and responsibility. *Philos Perspect* 18(1):189–208. <https://doi.org/10.1111/j.1520-8583.2004.00025.x>
- Maitra I, McGowan MK (2012) Introduction and overview. In: Maitra I, McGowan MK (eds) *Speech and harm*. Oxford University Press, pp 1–23. <https://doi.org/10.1093/acprof:oso/9780199236282.001.0001>
- Matsuda MJ (2018) *Words that wound: critical race theory, assaultive speech, and the first amendment*. Routledge
- McGowan MK (2018) Responding to harmful speech: the more speech response, counter speech, and the complexity of language use. In: Johnson CR (ed) *Voicing Dissent*. Routledge, pp 182–199. <https://lccn.loc.gov/2017061301>. Accessed 10 Nov 2023
- Monteith MJ, Deneen NE, Tooman GD (1996) The effect of social norm activation on the expression of opinions concerning gay men and blacks. *Basic Appl Soc Psychol* 18(3):267–288. https://doi.org/10.1207/s15324834basp1803_2
- Nielsen LB (2012) Power in public. In: Maitra I, McGowan MK (eds) *Speech and harm*. Oxford University Press, pp 148–173. <https://doi.org/10.1093/acprof:oso/9780199236282.003.0007>
- Opp KD (2001) How do norms emerge? An outline of a theory. *Mind Soc* 2(1):101–128. <https://doi.org/10.1007/bf02512077>
- Opp KD (2002) When do norms emerge by human design and when by the unintended consequences of human action? The example of the no-smoking norm. *Ration Soc* 14(2):131–158. <https://doi.org/10.1177/1043463102014002001>
- Parkinson B, Manstead ASR (1993) Making sense of emotion in stories and social life. *Cogn Emot* 7(3–4):295–323. <https://doi.org/10.1080/02699939308409191>
- Perry B, Alvi S (2012) We are all vulnerable. *Int Rev Victimol* 18(1):57–71. <https://doi.org/10.1177/0269758011422475>
- Petersen TS (2016) No offense! On the offense principle and some new challenges. *Crim Law Philos* 10(2):355–365. <https://doi.org/10.1007/s11572-014-9333-2>

- Przepiorka W, Szekely A, Andrighetto G, Diekmann A, Tummolini L (2022) How norms emerge from conventions (and change). *Socius*, 8. <https://doi.org/10.1177/23780231221124556>
- Rosenfeld M (2003) Hate speech in constitutional jurisprudence: a comparative analysis. *Cardozo L Rev* 24(4):1523–1567
- Rovira A, Southern R, Swapp D, Campbell C, Zhang JJ, Levine M, Slater M (2021) Bystander affiliation influences intervention behavior: a virtual reality study. *SAGE Open*, 11(3). <https://doi.org/10.1177/21582440211040076>
- Schein C, Gray K (2018) The theory of dyadic morality: reinventing moral judgment by redefining harm. *Personal Soc Psychol Rev* 22(1):32–70. <https://doi.org/10.1177/1088868317698288>
- Soral W, Bilewicz M, Winiewski M (2018) Exposure to hate speech increases prejudice through desensitization. *Aggress Behav* 44(2):136–146. <https://doi.org/10.1002/ab.21737>
- Sullivan GM, Artino AR (2013) Analyzing and interpreting data from likert-type scales. *J Grad Med Educ* 5(4):541–542. <https://doi.org/10.4300/JGME-5-4-18>
- Swim JK, Hyers LL (1999) Excuse me—what did you just say?!: Women’s public and private responses to sexist remarks. *J Exp Soc Psychol* 35(1):68–88. <https://doi.org/10.1006/jesp.1998.1370>
- Tunçgenç B, El Zein M, Sulik J, Newson M, Zhao Y, Dezechache G, Deroy O (2021) Social influence matters: we follow pandemic guidelines most when our close circle does. *Br J Psychol* 112(3):763–780. <https://doi.org/10.1111/bjop.12491>
- Waldron J (2012) *The harm in hate speech*. Harvard University Press. <https://www.jstor.org/stable/j.ctt2jbrjd>
- Walters MA (2014) The harms of hate crime: from structural disadvantage to individual identity. In: *hate crime and restorative justice: exploring causes, repairing harms*. Oxford University Press, pp 62–90 <https://doi.org/10.1093/acprof:oso/9780199684496.003.0003>
- Witek M (2013) How to establish authority with words. *Logic. Methodol Philos Sci Wars Univ* 2(2011):145–157
- Wong RYM, Cheung CMK, Xiao B, Thatcher JB (2021) Standing up or standing by: understanding Bystanders’ proactive reporting responses to social media harassment. *Inf Syst Res* 32(2):561–581. <https://doi.org/10.1287/isre.2020.0983>
- Zapata J, Deroy O (2023) Ordinary citizens are more severe towards verbal than nonverbal hate-motivated incidents with identical consequences. *Sci Rep* 13(1):1–14. <https://doi.org/10.1038/s41598-023-33892-8>
- Zitek EM, Hebl MR (2007) The role of social norm clarity in the influenced expression of prejudice over time. *J Exp Soc Psychol* 43(6):867–876. <https://doi.org/10.1016/j.jesp.2006.10.010>

Acknowledgements

JZ and CW were supported by the Mentoring Programme from the Faculty of Philosophy, Philosophy of Science and Religious Studies at Ludwig-Maximilians-Universität München. The funders had no role in study design, data collection and analysis, the decision to publish or the preparation of the manuscript. The authors thank Dr Ivar Hannikainen for helpful comments on previous versions of the experimental design.

Author contributions

JZ: conceptualization, methodology, investigation, writing - original draft, project administration, funding acquisition; CW: formal analysis, investigation, data curation,

resources; JS: formal analysis, visualisation, data curation, writing - review & editing; OD: methodology, writing - review & editing, supervision and funding acquisition.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

JS was a member of the Editorial Board of this journal at the time of acceptance for publication. The manuscript was assessed in line with the journal’s standard editorial processes. JZ, CW, and OD declare no potential conflict of interest.

Ethical approval

This study was performed in line with the principles of the Declaration of Helsinki. The Ethics Committee of the Faculty of Philosophy, Philosophy of Science and Religious Studies at the Ludwig-Maximilians-Universität München approved the protocol for this study (ID-Number 131874/ February 10, 2022).

Informed consent

All participants provided informed consent before taking part in the study. They received relevant information about the research aim, procedure, duration, and compensation. Furthermore, we informed them that although some visual scenes could be distressful, participating in the experiment would involve no other expected risks and that they could withdraw from it at any time without further consequences.

Additional information

Correspondence and requests for materials should be addressed to Jimena Zapata.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024