The breakdown of antiracist norms: A natural experiment on hate speech after terrorist attacks

Amalia Álvarez-Benjumea[a,1] and Fabian Winter[a]

[a]Max-Planck Research Group "Mechanisms of Normative Change," Max-Planck-Institute for Research on Collective Goods, 53113 Bonn, Germany

Edited by Douglas S. Massey, Princeton University, Princeton, NJ, and approved August 3, 2020 (received for review April 24, 2020)

Terrorist attacks often fuel online hate and increase the expression of xenophobic and antiminority messages. Previous research has focused on the impact of terrorist attacks on prejudiced attitudes toward groups linked to the perpetrators as the cause of this increase. We argue that social norms can contain the expression of prejudice after the attacks. We report the results of a combination of a natural and a laboratory-in-the-field (lab-in-the-field) experiment in which we exploit data collected about the occurrence of two consecutive Islamist terrorist attacks in Germany, the Würzburg and Ansbach attacks, in July 2016. The experiment compares the effect of the terrorist attacks in hate speech toward refugees in contexts where a descriptive norm against the use of hate speech is evidently in place to contexts in which the norm is ambiguous because participants observe antiminority comments. Hate toward refugees, but not toward other minority groups, increased as a result of the attacks only in the absence of a strong norm. These results imply that attitudinal changes due to terrorist attacks are more likely to be voiced if norms erode.

terrorist attacks | social norms | online hate | prejudice | refugees

On 18 July 2016, a 17-y old armed with an ax attacked passengers on board a train heading to Würzburg in the southern part of Germany. Six days later, on 24 July, another attacker injured several people and killed himself when he detonated a backpack bomb in Ansbach, near Nuremberg, in the first Islamist terrorist suicide attack in Germany. Both attacks were later claimed by the Islamic State (IS). The two consecutive terrorist attacks hit Germany at the peak of the so-called "European refugee crisis." During this period, civil wars in Syria and Iraq caused a massive displacement of people fleeing war and political instability, pushing large numbers of refugees to the surrounding countries and Europe. The situation fueled an already heated public discussion on German policies on immigration.

After terrorist attacks, hatred often follows suit (1–5). The effect is particularly noticeable when the attacker is characterized as a member of a social or religious minority, as exemplified by the wave of anti-Muslim hate crimes that followed the 9/11 terrorist attacks (4–6), the increase in violence against refugees linked to Islamist attacks in Germany (7), or the escalation of hate speech on Twitter after an Islamist attack in the United Kingdom (8). More generally, formal and informal norms of "civic behavior" seem to erode after such attacks, and behavior that was not acceptable before becomes more frequent in the aftermath.

We explain the erosion of civic behavior by focusing on one of the most immediate public reactions to terrorist attacks that can usually be observed in social media: The expression of prejudice gains traction in online environments (1, 2). We refer to this as hate speech, which is speech intended to promote hatred on the basis of race, religion, ethnicity, or sexual orientation (9). Widespread hate speech may cause anger, frustration, or resignation (10) and pushes people out of the public debate (11), thus harming the free exchange of opinions and ideas in the long run. Therefore, understanding how dramatic events might affect online hate speech is important to prevent a toxic online environment and promote open conversations.

As of now, however, little is known about the mechanisms causing this increase. It is well established through observational studies that terrorist attacks have a profound impact on xenophobic attitudes (12–14), particularly those that occurred in national territory (15), which has led many scholars to assume that the rise in online hate results from the change in attitudes (1, 6). The attitudinal change argument states that terrorist attacks increase xenophobic attitudes and antiimmigrant sentiment because people perceive terrorist attacks carried out by out-group members as intergroup threat (7, 15, 16). This leads to an increase in prejudice (17, 18) and results in an increase in hate speech as a direct consequence of the change in attitudes.

Focusing solely on a change in individual attitudes misses a crucial point: Hate speech is a communicative act and, as such, it is regulated by social norms. Social norms play a decisive role in containing the public expression of prejudice (19–21), such as xenophobic, racist, and discriminatory remarks. Previous research found that norms have a direct effect on behavior that is independent of individual attitudes (22). Because of the independent effects of norms and individual attitudes, an increment of antiminority sentiment after terrorist attacks or a legitimization of preexisting prejudice will materialize in more hateful comments only if the social norm allows it.

We empirically test the role of social norms in containing the expression of online hate after terrorist attacks using a combination of a natural experiment and a laboratory-in-the-field (lab-in-the-field) experiment. Between two waves of data collection in a lab-in-the-field experiment on hate speech in an experimental online discussion forum, the Würzburg and Ansbach terrorist attacks took place consecutively in Germany. We analyze the impact of these terrorist attacks on hate speech in

Significance

Surges in hateful and xenophobic content online are often found after terrorist attacks. We find that this effect is highly dependent on the local context and the respective social norms. Prejudiced attitudes are likely to be voiced only if the perceived social acceptability of expressing prejudice increases. Since antihate norms play an important role in containing the expression of prejudice, understanding how terrorist attacks may impact the strength of the social norm is essential to understanding societal responses to terrorist attacks.

Author contributions: A.A.-B. and F.W. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: alvarezbenjumea@coll.mpg.de.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2007977117/-/DCSupplemental.

First published September 1, 2020.

an online forum purpose built to investigate hate speech toward refugees and gender issues. Since there are very few formal rules in many online contexts, social norms play a crucial role in these domains (20, 21), making these forums a perfect setting to test our hypothesis.

Our main argument is that people will be more likely to express prejudiced attitudes after the attacks if the validity of antihate speech norms is challenged, but they will refrain from doing so if they perceive the norm to remain strong. First, we need to assess that the terrorist attacks had an impact on hate speech against refugees. The online forum features comments on gender rights as well as on refugees. We selected comments on gender rights as a comparison group. Because this category is completely unrelated to the attacks, comments in the forum should not be affected by them.

Second, we test whether the social norm against the public expression of hate limits the expression of antirefugee sentiment. We compare conditions in which the norm is left ambiguous to conditions in which the norm is presented as strong. The strength of the norm is signaled by what others do. If many others engage in the expression of prejudice, other participants may logically assume that they approve of this behavior (23, 24). If they never engage in it, participants might expect that they disapprove. The descriptive norm changes expectations about how appropriate the expression of prejudice is in the specific context (25–27). Our experimental conditions thus vary in the composition of the comments the participants observe. We compare contexts where a descriptive norm against the use of hate speech is strong because no hateful comments can be observed to contexts in which the norm is deliberately left ambiguous because xenophobic comments are observed. We expect the descriptive norm to have a deterrent effect on the expression of prejudice.

Concerning the attitudinal change argument, it is important to stress that randomization into conditions ensures that any increases in individual prejudice or in the willingness to express it after the attacks are similar across conditions. Potential differences in expressed prejudice can therefore be solely attributed to the effect of the social norm.

Experimental Design: A Combination of a Lab-in-the-Field Experiment and a Natural Experiment

We collected our data in a purpose-built online forum. The forum was designed as a realistic yet carefully controlled environment where participants were invited to discuss selected social topics (the chronological steps for constructing the online forum are described in *SI Appendix*, Fig. S1). The forum featured discussions on two social topics: gender rights and refugees. Public opinion linked the surge in Islamist terrorism to the recent increase of refugees and fed a narrative around refugees as threatening security and Western values (28) (for further discussion on the political situation see *SI Appendix*, section 1). We refer to the attacks as the treatment in the experimental jargon. Hence, we consider comments on refugees as the treated group and use comments on gender rights as the control group. Due to the unforeseen circumstances of the data collection, the requirement for ethical approval was waived after data collection by the Ethics Commission of the University of Bonn.

Participants in the Experiment. A total of 139 different participants for our experiment before the terrorist attacks and 135 after the terrorist attacks were recruited via a crowdsourcing internet marketplace, resulting in a total of 2,133 comments. Demographic information on the general characteristics of the workforce for reference is in *SI Appendix*, Table S3. The experiment was conducted entirely in German and the sample was strictly restricted to residents in Germany. Participants voluntarily registered for the experiment via an online marketplace. We randomized participants into experimental conditions. Table 1 shows the number of comments collected by time of data collection (e.g., before or after the terrorist attacks), experimental condition, and topic.

Participating in the Experimental Forum. At the beginning of the experiment, participants provided informed consent to participate in the study and were given a user name and an avatar (*SI Appendix*, section 3 and Fig. S2). They were asked to join the conversations and instructed to leave comments about pictures that portrayed different social topics (see *SI Appendix*, section 4 for the instructions). Once the experiment started, every participant was consecutively presented with the discussions and asked to leave a comment at the bottom of each thread (see Fig. 1 for a screenshot of the online forum). Participants could see only the comments we had previously selected to create the different conditions. This ensures that individual observations are independent and increases internal validity (29) (for a discussion on the validity of the design, see *SI Appendix*, section 7). Each participant was required to leave a comment on each forum page, with a total of eight comments per participant. No further identifying information was collected from the participants.

Experimental Conditions. Participants in the forum were randomized into three different experimental conditions: a no-norm, a weak-norm, and a strong-norm condition. Each condition consisted of a different mix of comments, from friendly language to actual transgressions of the antihate norm. *SI Appendix*, Table S1 shows a summary of the forum content in the different conditions. In the no-norm condition no specific norm is signaled. The no-norm condition featured a mix of comments, including hate speech examples, that did not signal any specific descriptive norm; the presented comments ranged from hostile to very positive. Therefore, the acceptability of using hate speech is ambiguous. In the weak-norm and strong-norm conditions, an antihate descriptive norm is highlighted with different strengths. In the weak-norm condition, we removed all hostile comments and thus biased the perception of how many others use hate speech. This created a behavioral regularity that signals the existence of a descriptive norm against the use of hate in the online forum. The strong-norm condition further emphasized the descriptive norm by showing only very positive comments toward the respective groups.

The aim of this study is to provide an experimental test of whether the voicing of changing opinions depends on the normative context: Negative attitudes translate into negative comments only if norms are challenged and the speaker observes a "negative" environment. In this case, we expect an increase in hateful comments toward refugees, but not in gender rights in the no-norm condition since gender norms are not challenged by the attack. The increase in hate comments against refugees should be reduced or nonexistent in the strong-norm condition since attitudes may have changed due to the attacks, but the descriptive norms against hate speech remained strong. If, on the contrary,

Table 1. Number of comments (N = 2,133) per time of data collection (before and after the terrorist attacks), experimental condition, and topic

	Before the attacks		After the attacks	
Condition	Refugees	Gender rights	Refugees	Gender rights
No norm	135	227	135	228
Weak norm	136	225	135	226
Strong norm	134	225	123	204
Total	405	677	393	658

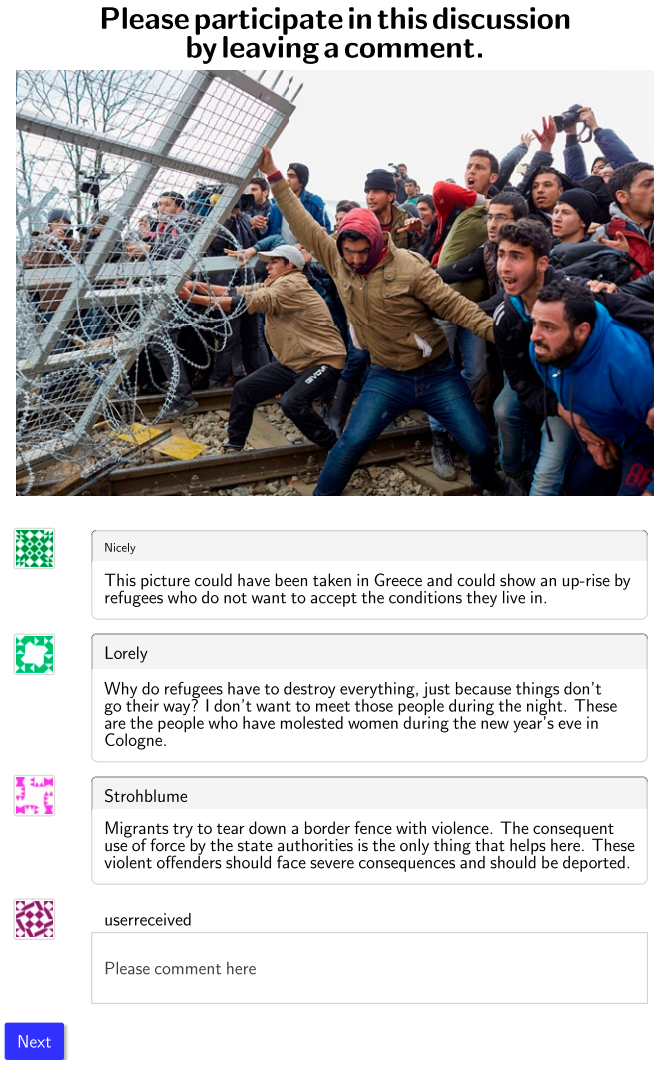

Fig. 1. Screenshot of the no-norm experimental condition (our translation; in German in the original). Image credit: Getty Images/Pierre Crom.

social norms have no effect, we would expect hate speech to increase similarly for all conditions and topics.

Construction of the Hate Score. Our main focus in this study is changes in hate speech, i.e., expressed prejudice toward minority groups displayed by the participants in their comments. We therefore constructed a hate score to measure change before and after the terrorist attacks and across the different conditions. To construct the hate score, we asked 577 external raters to rate how prejudiced the comments were several weeks after the data collection in both waves. The raters were recruited from the same population as the participants (*SI Appendix*, Table S3). Every rater was assigned a set of about 30 randomly chosen comments and then an average score was computed for each comment (in *SI Appendix*, Fig. S5, we conduct analyses of the ratings). The rating task was completed online. We asked the raters to rate the comment on the following scale: "Is the comment friendly or hostile toward the group represented in the picture? (Give a number from 1 to 9 where 1 means very friendly and 9 means very hostile)." Comments with lower scores, e.g., 1 to 4, are therefore affable, with a cordial language, and often express a strong-norm opinion. On the other side of the spectrum, high scores such as 8 or 9 generally imply abusive language, e.g., "I cannot stand gay people. They should have a psychiatric exam" or the use of hate terms, e.g., "They can continue walking away from Europe. They are not just war refugees, 90 per cent are nothing but social parasites who do whatever they want here" (emphasis added). *SI Appendix*, section 5 contains examples of comments and their classification. *SI Appendix*, Table S4 gives an overview about the descriptive statistics of the mean hate score.

Results

This section provides the main results of our experiment. *SI Appendix* provides further details regarding the statistical approach (*SI Appendix*, section 8) and the statistical models (*SI Appendix*, section 9), as well as auxiliary analyses (*SI Appendix*, section 10).

Hate toward Refugees but Not toward Other Groups Soared after the Terrorist Attacks. We first check whether the attacks had an impact on hate speech against refugees in our online forum. For this purpose, we analyze the effect of the terrorist attacks on hate speech in the no-norm condition (N = 725 comments). As described before, the no-norm condition features a mix of comments containing examples of antiminority and xenophobic comments; therefore, the descriptive norm against the expression of hate in this condition is ambiguous.

SI Appendix, Table S5 provides the results from random-effects multilevel regression models, which take account of the nested design of comments in participants. The results are also depicted in Fig. 2. The mean hate score increases by 0.56 points following the attacks ($P = 0.004$; see model 1 in *SI Appendix*, Table S5 and Fig. 2*A*). To give an intuition about what these coefficients mean, we can look at how changes in the score are translated into changes in hostility in the comments. A change in one score point can be very noticeable when comparing two comments from a thread on the same topic. A comment with a score of 6 reads "Get rid of the funny get-up . . .We are in Europe, or more precisely Germany. Whoever wants to live here has to adapt. Multiculturalism, sure, but not that much." A comment in the same picture, but with a score of 7 reads "That's a very special bird, a *black barn owl* . . ." (in reference to a woman wearing a burqa).

We next compare the postattacks change in hate speech against refugees to postattacks changes in gender rights to isolate the effect of the terrorist attacks. We assume that the terrorist

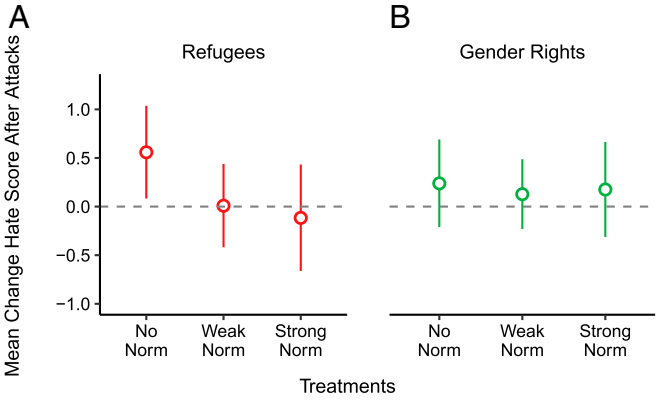

Fig. 2. (*A* and *B*) Postattacks mean change estimates of hate score in comments on refugees (*A*) and comments on gender rights (*B*) for the different levels of the descriptive norm: no norm (N = 257), weak norm (N = 271), and strong norm (N = 270). Estimates account for the nested structure of the data with a random intercept multilevel linear regression with an effect for participants. The regressions are conducted separately for each treatment. The dashed line represents no change in hate after the attacks. Error bars represent the 95% confidence interval of the estimates.

attacks would not have an effect on hate toward unrelated topics. This increase is certainly not found in comments discussing gender rights. After the attacks, comments on gender rights became 0.23 points more hostile, but this effect is not statistically significant ($P = 0.28$) and substantially smaller than the 0.59 points increase for postattacks comments about refugees. Also the difference in difference of 0.36 points is only marginally significant ($P = 0.063$; see the interaction term After Attacks × Refugees in *SI Appendix*, Table S5). The results support the finding that the attacks increased hate speech against refugees in the online forum.

Social Norms Contain the Expression of Hate Speech after Terrorist Attacks. Fig. 2 depicts the postattacks changes in hate score for comments on refugees. Markers represent regression coefficients and lines the 95% confidence intervals of the estimate. The regressions are conducted separately for each treatment and test whether, for each group, the estimated coefficient is significantly different from zero.

We predict that hate speech against refugees does not increase in the online forum in the presence of a descriptive norm against its expression. We test this prediction by comparing the effect of the terrorist attacks on comments in the no-norm condition to their effect in the weak-norm and strong-norm forums. Following our argument, we expect the descriptive norm to contain the expression of xenophobic and antiimmigrant comments. Fig. 2 depicts the postattacks average change in hate score in the three experimental conditions for comments on refugees. There is no pre- and postattacks difference in hate in either the weak or the strong-norm condition in comments about refugees. Model 3 in *SI Appendix*, Table S5 provides corroborating results from random-effects multilevel regression models that the postattacks increase of 0.56 points in the no-norm condition is offset in both the weak-norm condition ($\beta = -0.55$, $P = 0.11$) and the strong-norm condition ($\beta = -0.68$, $P = 0.054$). If the participants are confronted with weak-norm and strong-norm comments only, the amount of prejudice they express is statistically indistinguishable before and after the attacks ($\beta = 0.014$, $P = 0.95$). This deterrent effect of the descriptive norm after the attacks is not found in topics unrelated to the attacks, i.e., gender rights (*SI Appendix*, Table S5).

If we restrict the analysis to postattacks comments, the mean hate score is significantly smaller in the weak-norm ($\beta = -0.48$, $P = 0.04$) and the strong-norm condition ($\beta = -0.62$, $P = 0.01$) compared to the no-norm condition, which means that the main effects of the experimental conditions became significant after the terrorist attacks. These findings are consistent with our claim that the increase in hate speech occurs only when the descriptive norms are ambiguous.

The descriptive norm prevented participants from expressing xenophobic and antiimmigrant opinions after the attacks in the weak-norm and strong-norm conditions. The increase in hate speech in the forum after the terrorist attacks therefore cannot be solely attributed to an increase in negative attitudes toward refugees. An increase in hate resulting exclusively from an increase in attitudes would have been consistent across all conditions.

Descriptive Norms Have the Greatest Effect on Extreme Comments. Finally, we show that the postterrorist attacks difference between the conditions is driven by the most extreme comments, as the largest differences between the conditions emerge in the 75th quantile of the distribution of the hate score and above. We estimate a quantile regression model to show how the magnitude of the effect of the terrorist attacks varies across the different percentiles of this distribution of the hate score. Fig. 3 depicts the postattacks change on comments about refugees for the quantiles 10th to 95th of the distribution of the hate

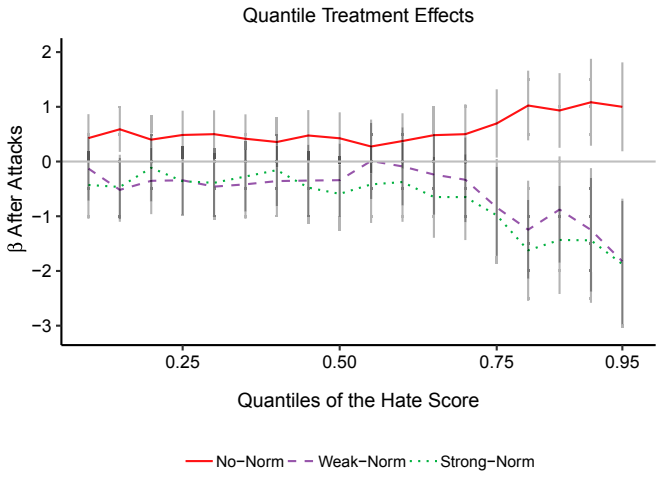

Fig. 3. The plot depicts the estimated model coefficients of the effect on terrorist attacks for quantiles 10th to 95th in comments on refugees for all levels of the descriptive norm: no norm, weak norm, and strong norm. The reference category is the preattacks distribution of the hate score in the no-norm condition. The gray vertical lines represent the confidence interval of the quantile regression coefficients for the effect of terrorist attacks with 95% confidence level.

score with the 95% confidence intervals. Each gray line represents the coefficient for the quantile indicated on the x axis. Each coefficient corresponds to the change in the τth quantile after the terrorist attacks compared to the hate score before the attacks in the no-norm condition. The results are shown for all three experimental conditions: no norm, weak norm, and strong norm.

The estimated gap between the three conditions tends to become more pronounced closer to the upper tail of the distribution of the hate score, particularly from the 75th quantile on. In the no-norm condition, the terrorist attacks increased hate speech along the entire distribution, but this effect is stronger in the higher or "hostile" quantiles of the distribution. Compared to the average change of 0.56 points in the score in the no-norm condition after the terrorist attacks, the 80th and the 90th quantiles have an estimated increase of 1.023 and 1.083 points, respectively ($P < 0.05$). At the 80th quantile comments are more than one point more hateful after the attacks (*SI Appendix*, Table S8). In the weak-and strong-norm conditions, the effect is also larger for highest quantiles but negative: More hateful comments became less frequent after the attacks. Overall, this suggests that the average changes after the attacks result from extremely hateful comments becoming more likely in the no-norm condition and less likely in the weak-norm and strong-norm condition after the attacks.

Discussion

Our study shows the importance of social norms for containing hate speech after terrorist attacks. We acknowledge that terrorist attacks, as previous empirical research found, might increase negative attitudes—and their expression—toward certain social groups. However, we argue that prejudice will be publicly voiced only if hate speech also becomes socially permissible. After the attacks, people search for cues in their environment on how to behave, and the behavior of others provides such cues in the form of descriptive social norms. Depending on these social cues, the prevalence of norm violations may increase, stay the same, or even decrease. Attitudinal change materializes in public transgressions of social norms only if the norms are challenged by the event. We apply our reasoning to explain the erosion of norms of civic conversations after terrorist attacks in an online context. Our study empirically confirms this finding and provides

additional evidence on how the expression of hate is regulated by social norms.

We find that hate speech toward refugees increased after the terrorist attacks only if participants could observe previous hate comments. Online hate against refugees increases after the attacks both compared with levels of hate against refugees before the attacks and relative to the increase toward gender rights. This increase is not found when the social norm is exogenously manipulated to remain strong. Under the assumption of proper randomization of individual attitudes into experimental conditions, this difference can be solely attributed to the normative context.

Furthermore, we show that the terrorist attacks radicalize the already hateful comments under certain conditions, but have only a small effect on positive or moderate ones. Extremely hateful comments become more frequent after the attacks when the norm is ambiguous, but less frequent in the conditions with a strong norm. Descriptive norms against hate speech thus act as a bulwark that prevents extremely prejudiced opinions from being voiced. On the positive side, the vast majority of people would not be converted into spreaders of hate, just because others do so. However, it also suggests that those who already hold negative beliefs about minorities seem to be encouraged to paint an even darker picture of immigration in Western societies.

Of course, this study paints a simplified picture of online communities in which participants are chosen randomly, interact only once, and are anonymous. Furthermore, we do not explore how different individuals may react differently, such as racist trolls or other instigators of antisocial behavior. Future research should consider repeated interaction, nonanonymity, or heterogeneous effects to expand the results. For the present purposes, the design suffices to show the importance of descriptive norms in shaping the reaction to terrorist attacks in online communities.

Our results suggest that supervisors of public discussions, either in the virtual domain or in the real world, may be well advised to implement measures to ensure a well-tempered atmosphere. Particularly after events leading to a broad discussion about the status of minorities, too many bad examples could lead to an erosion of norms and embolden prejudiced citizens. As an unintended consequence, this may lead to the (self-)exclusion of marginalized groups from the discussion and in the worst case to a breakdown of the whole debate.

Data Availability. Anonymized comma-separated value (csv) code have been deposited in Open Science Framework (DOI 10.17605/OSF.IO/3S5PK) (30).

ACKNOWLEDGMENTS. We thank Emad Bahrami Rad for the programming assistance. Valuable feedback in earlier drafts of this article was provided by Nan Zhang, Rima Maria Rahal, Delia Baldasarri, and Michael Mäs; the members of the Max Planck Institute for Research on Collective Goods; the members of the Norms and Networks cluster at the University of Groningen; and conference participants at the 12th Annual Meeting of the International Network of Analytical Sociologists (INAS 2019), the 6th International Meeting on Experimental and Behavioral Social Sciences (IMEBESS 2019), the 3rd Cultural Transmission and Social Norms workshop (CTSN 2018), and the seminar of the Democracy, Elections, and Citizenship (DEC) research group at the Autonomous University of Barcelona. Financial support of the Max-Planck-Society for the Max-Planck Research Group "Mechanisms of Normative Change" is gratefully acknowledged.

1. I. Awan, I. Zempi, 'I will blow your face OFF'–Virtual and physical world anti-Muslim hate crime. *Br. J. Criminol.* **57**, 362–380 (2017).
2. P. Burnap *et al.*, Tweeting the terror: Modelling the social media reaction to the Woolwich terrorist attack. *Soc. Network Anal. Min.* **4**, 206 (2014).
3. R. D. King, G. M. Sutton, High times for hate crimes: Explaining the temporal clustering of hate-motivated offending. *Criminology* **51**, 871–894 (2013).
4. I. Disha, J. C. Cavendish, R. D. King, Historical events and spaces of hate: Hate crimes against Arabs and Muslims in post-9/11 America. *Soc. Probl.* **58**, 21–46 (2011).
5. B. D. Byers, J. A. Jones, The impact of the terrorist attacks of 9/11 on anti-Islamic hate crime. *J. Ethn. Crim. Justice* **5**, 43–56 (2007).
6. E. Hanes, S. Machin, Hate crime in the wake of terror attacks: Evidence from 7/7 and 9/11. *J. Contemp. Crim. Justice* **30**, 247–267 (2014).
7. S. Jäckle, P. D. König, Threatening events and anti-refugee violence: An empirical analysis in the wake of the refugee crisis during the years 2015 and 2016 in Germany. *Eur. Sociol. Rev.* **34**, 728–743 (2018).
8. M. L. Williams, P. Burnap, Cyberhate on social media in the aftermath of Woolwich: A case study in computational criminology and big data. *Br. J. Criminol.* **56**, 211–238 (2015).
9. I. Gagliardone, D. Gal, T. Alves, G. Martinez, *Countering Online Hate Speech* (UNESCO Publishing, 2015).
10. S. Hinduja, J. W. Patchin, Offline consequences of online victimization: School violence and delinquency. *J. Sch. Violence* **6**, 89–112 (2007).
11. C. West, "Words that silence? Freedom of expression and racist hate speech" in *Speech and Harm: Controversies over Free Speech*, I. Maitra, M. K. McGowan, Eds. (Oxford University Press, Oxford, 2012), pp. 222–248.
12. M. Walters, R. Brown, S. Wiedlitzka, Causes and motivations of hate crime. https://ssrn.com/abstract=2918883. Accessed 12 February 2020.
13. A. Echebarria-Echabe, E. Fernández-Guede, Effects of terrorism on attitudes and ideological orientation. *Eur. J. Soc. Psychol.* **36**, 259–265 (2006).
14. H. G. Boomgaarden, C. H. de Vreese, Dramatic real-world events and public opinion dynamics: Media coverage and its impact on public reactions to an assassination. *Int. J. Public Opin. Res.* **19**, 354–366 (2007).
15. J. Legewie, Terrorist events and attitudes toward immigrants: A natural experiment. *Am. J. Sociol.* **118**, 1199–1245 (2013).
16. B. Doosje, A. Zimmermann, B. Küpper, A. Zick, R. Meertens, Terrorist threat and perceived Islamic support for terrorist attacks as predictors of personal and institutional out-group discrimination and support for anti-immigration policies–evidence from 9 European countries. *Rev. Int. Psychol. Soc.* **22**, 203–233 (2009).
17. B. M. Riek, E. W. Mania, S. L. Gaertner, Intergroup threat and outgroup attitudes: A meta-analytic review. *Pers. Soc. Psychol. Rev.* **10**, 336–353 (2006).
18. I. Fritsche, E. Jonas, T. Kessler, Collective reactions to threat: Implications for intergroup conflict and for solving societal crises. *Soc. Issues Policy Rev.* **5**, 101–136 (2011).
19. F. A. Blanchard, C. S. Crandall, J. C. Brigham, L. A. Vaughn, Condemning and condoning racism: A social context approach to interracial settings. *J. Appl. Psychol.* **79**, 993–997 (1994).
20. A. Álvarez-Benjumea, F. Winter, Normative change and culture of hate: An experiment in online environments. *Eur. Sociol. Rev.* **34**, 223–237 (2018).
21. K. Munger, Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Polit. Behav.* **39**, 629–649 (2017).
22. E. L. Paluck, Reducing intergroup prejudice and conflict using the media: A field experiment in Rwanda. *J. Pers. Soc. Psychol.* **96**, 574 (2009).
23. C. Horne, S. Mollborn, Norms: An integrated framework. *Annu. Rev. Sociol.* **46**, 467–487 (2020).
24. K. D. Opp, "What is always becoming what ought to be.": How political action generates a participation norm. *Eur. Sociol. Rev.* **20**, 13–29 (2004).
25. M. E. Tankard, E. L. Paluck, Norm perception as a vehicle for social change. *Soc. Issues Policy Rev.* **10**, 181–211 (2016).
26. C. Bicchieri, *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms* (Oxford University Press, New York, NY, 2016).
27. Bicchieri, C., & Dimant, E. Nudging with care: The risks and benefits of social information. *Public Choice*, 1–22 (2019).
28. E. Greussing, H. G. Boomgaarden, Shifting the refugee narrative? An automated frame analysis of Europe's 2015 refugee crisis. *J. Ethnic Migrat. Stud.* **43**, 1749–1774 (2017).
29. H. Rauhut, F. Winter, "On the Validity of Laboratory Research in the Political and Social Sciences: The Example of Crime and Punishment" in *Experimental Political Science*, B. Kittel, W. J. Luhan, R. B. Morton, Eds. (Palgrave Macmillan, London, 2012), pp. 209–232.
30. A. Álvarez-Benjumea, F. Winter, Supplementary materials the breakdown of anti racist norms: A natural experiment on hate speech after terrorist attacks. Open Science Framework. https://osf.io/3s5pk. Deposited 10 July 2020.