Empathy-based counterspeech can reduce racist hate speech in a social media field experiment

Dominik Hangartner[a,b,1], Gloria Gennaro[a,b], Sary Alasiri[a], Nicholas Bahrich[a], Alexandra Bornhoft[a], Joseph Boucher[a], Buket Buse Demirci[a], Laurenz Derksen[a,b], Aldo Hall[a], Matthias Jochum[a], Maria Murias Munoz[a], Marc Richter[a], Franziska Vogel[a], Salomé Wittwer[a], Felix Wüthrich[a], Fabrizio Gilardi[c], and Karsten Donnay[c]

[a]Center for Comparative and International Studies, Eidgenössische Technische Hochschule Zurich, 8092 Zurich, Switzerland; [b]Immigration Policy Lab, Eidgenössische Technische Hochschule Zurich, 8092 Zurich, Switzerland; and [c]Department of Political Science, University of Zurich, 8050 Zurich, Switzerland

Edited by Margaret Levi, Department of Political Science, Stanford University, Stanford, CA; received September 3, 2021; accepted November 2, 2021

Despite heightened awareness of the detrimental impact of hate speech on social media platforms on affected communities and public discourse, there is little consensus on approaches to mitigate it. While content moderation—either by governments or social media companies—can curb online hostility, such policies may suppress valuable as well as illicit speech and might disperse rather than reduce hate speech. As an alternative strategy, an increasing number of international and nongovernmental organizations (I/NGOs) are employing counterspeech to confront and reduce online hate speech. Despite their growing popularity, there is scant experimental evidence on the effectiveness and design of counterspeech strategies (in the public domain). Modeling our interventions on current I/NGO practice, we randomly assign English-speaking Twitter users who have sent messages containing xenophobic (or racist) hate speech to one of three counterspeech strategies—empathy, warning of consequences, and humor—or a control group. Our intention-to-treat analysis of 1,350 Twitter users shows that empathy-based counterspeech messages can increase the retrospective deletion of xenophobic hate speech by 0.2 SD and reduce the prospective creation of xenophobic hate speech over a 4-wk follow-up period by 0.1 SD. We find, however, no consistent effects for strategies using humor or warning of consequences. Together, these results advance our understanding of the central role of empathy in reducing exclusionary behavior and inform the design of future counterspeech interventions.

hate speech | social media | counterspeech | field experiment

There is increasing awareness among national and international policy makers, civil society organizations, and social media companies that online hate speech negatively impacts affected communities, which are disproportionately women, teenagers, LGTBQIA people, and minority ethnic groups (1). While hate speech—understood as hostile language against a group or person because of their actual or perceived innate characteristic (1)—is not a recent phenomenon, the rise of online forums provides perpetrators with higher visibility and allows for more direct targeting of their victims (2). A growing literature researches the "off-line" consequences of online hate speech and documents how it may harm mental and physical health, and trigger violence (3, 4). In addition, hate speech may polarize public opinion and hurt political discourse, with detrimental consequences for democracies (2).

The rise of online hate speech has been accompanied by efforts to curb it, including content moderation by governments and social media companies. Since content moderation may suppress valuable as well as illicit speech and disperse rather than reduce hate speech, international organizations and civil society organizations are increasingly turning to counterspeech as a strategy to confront hate speech (5). Unlike content moderation, counterspeech does not seek to suppress free expression, but instead promises to reduce hate through persuasion of the perpetrator (6). Despite their growing use, experimental evidence on the effectiveness and design of counterspeech interventions available to the public (rather than proprietary intellectual property of social media companies) is very limited.

Previous studies mostly focused on randomizing the characteristics of the counterspeaker and provided important insights on how social status moderates the effectiveness of counterspeech (7–9). In contrast, our field experiment holds the characteristics of the counterspeaker fixed but varies the content of the message. This focus on the content of counterspeech strategies increases the ecological validity and applicability of our findings for anti–hate speech campaigns. Building on existing categorizations of counterspeech, we focus on three of the most widely applicable and commonly used counterspeech strategies (10): humor, warning of consequences, and inducing empathy. Humor (and memes) is intended to shift and deescalate the dynamics of communication. Warning of consequences reminds the hate speech sender that family and acquaintances can also observe her public messages. Empathy seeks to humanize the victim and remind the sender that people can be hurt by her behavior.

To provide causal evidence on the effects of the three counterspeech strategies, we designed a field experiment focusing on reducing xenophobic and racist hate speech on the social media platform Twitter. As discussed in detail in *Materials and Methods*, we used a combination of dictionary-based approaches, sentiment analysis, and manual annotation to select English-speaking tweets containing xenophobic (targeting immigrants) and/or racist (targeting ethnic or racial minorities) content. Each of a total of $N = 1,350$ users who sent these tweets were randomly assigned (20% probability) to one of the three, one-shot, counterspeech interventions, or the control arm (40% probability), which received no intervention. For treated users, counterspeech was applied within 24 h of publication of the xenophobic tweet by a neutrally designed, human-controlled "bot." Fig. 1 provides examples of a bot and counterspeech messages. To comprehensively assess the counterspeech effects, we collected information on the retrospective deletion of past xenophobic tweets (deletion of the original xenophobic tweet and share of deleted tweets among the last 1,000 tweets sent before treatment), the future creation of xenophobic hate speech over a 4-wk follow-up period (number of xenophobic tweets, total number of tweets, and ratio of the two measures), and the average negative sentiment of all tweets in the follow-up period [measured using the Vader compound score (11)]. The experimental design and data collection were preregistered (see *Materials and Methods*).

Author contributions: D.H., S.A., N.B., A.B., J.B., B.B.D., L.D., A.H., M.J., M.M.M., M.R., F.V., S.W., F.W., F.G., and K.D. designed research; D.H., G.G., S.A., N.B., A.B., J.B., B.B.D., L.D., A.H., M.J., M.M.M., M.R., F.V., S.W., and F.W. performed research; D.H., G.G., L.D., S.W., and F.W. analyzed data; and D.H., G.G., F.G., and K.D. wrote the paper.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: dominik.hangartner@gess.ethz.ch.

Published December 6, 2021.

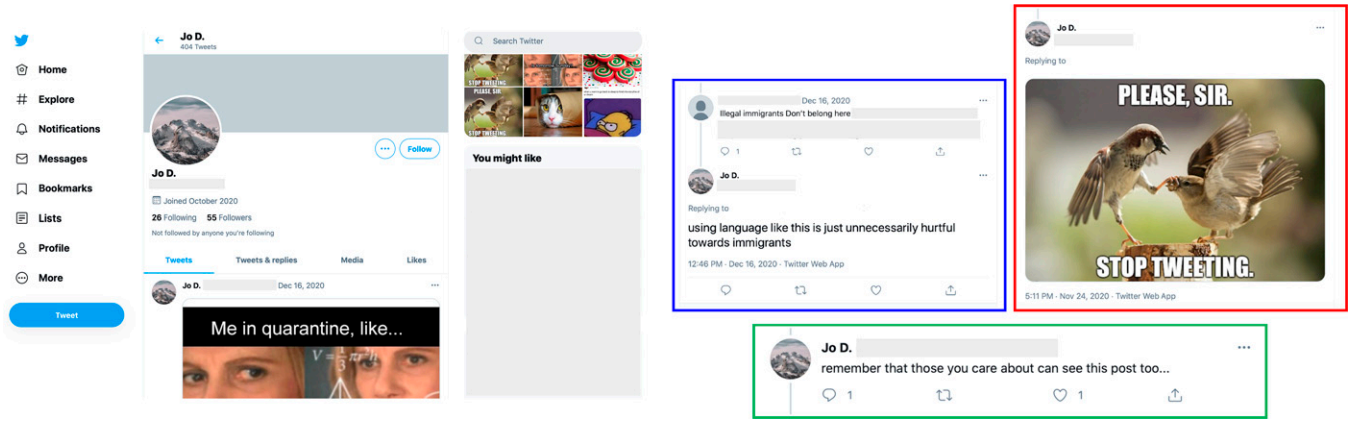

Fig. 1. Examples of a human-controlled "bot" account (*Left*) and counterspeech messages (*Right*) using warning of consequences (green box), humor (red box), and empathy-based (blue box) strategies. The hate tweet and Twitter handles are obscured to protect users' anonymity.

Results

To deal with a small number of tweets (5.5% of the sample) that the sender deleted after we assigned but before we applied the treatment, the intention-to-treat analysis leverages the randomized assignment of the treatment (instead of the treatment actually received). For each of the six outcomes and three treatment conditions (with the control group serving as reference category), we estimate separate ordinary least squares (OLS) regressions with robust SEs. These regressions also control for pretreatment covariates selected using Lasso-based postdouble selection (12). To account for multiple hypothesis testing, we also provide Benjamini–Hochberg (BH) P values, adjusted for each treatment arm, along with unadjusted estimates.

Fig. 2 shows the causal effects of the three counterspeech interventions on hate speech creation (Fig. 2, *Left*) and hate speech deletion, as well as overall "Vader" negativity (Fig. 2, *Right*). Compared to the control group, we find small but consistent effects suggesting that empathy-based counterspeech (blue estimates) reduces the volume of xenophobic tweets in the 4 wk after treatment by 0.10 SD ($\mathrm{SE} = 0.054$, $P = 0.056$, and $P_{BH} = 0.112$; all P values are from two-sided tests), and the total volume of tweets by 0.13 SD ($\mathrm{SE} = 0.065$, $P = 0.039$, and $P_{BH} = 0.112$). Furthermore, the empathy treatment increases the propensity to delete the original xenophobic tweet by 0.21 SD ($\mathrm{SE} = 0.081$, $P = 0.009$, and $P_{BH} = 0.053$). Converted to tweets, these estimates imply that users assigned to the empathy treatment sent, on average, 1.3 fewer xenophobic tweets and 91.6 fewer total tweets, and were 8.4 percentage points more likely to delete the original xenophobic tweet. Contrasting our preregistered hypotheses, we find no significant empathy effects on the remaining three outcomes, and generally smaller and not significant effects for counterspeech interventions based on humor (red estimates) and warning of consequences (green estimates) across all outcomes. This interpretation is supported by an F test that rejects the joint null hypothesis of no effect across outcomes for the empathy treatment ($P = 0.017$), but not for the humor ($P = 0.285$) and warning of consequences ($P = 0.633$) treatments (seemingly unrelated regression [SUR]; see *Materials and Methods*).

Discussion

Counterspeech is a widespread strategy to confront hate speech online. Our randomized field experiment shows that, of the

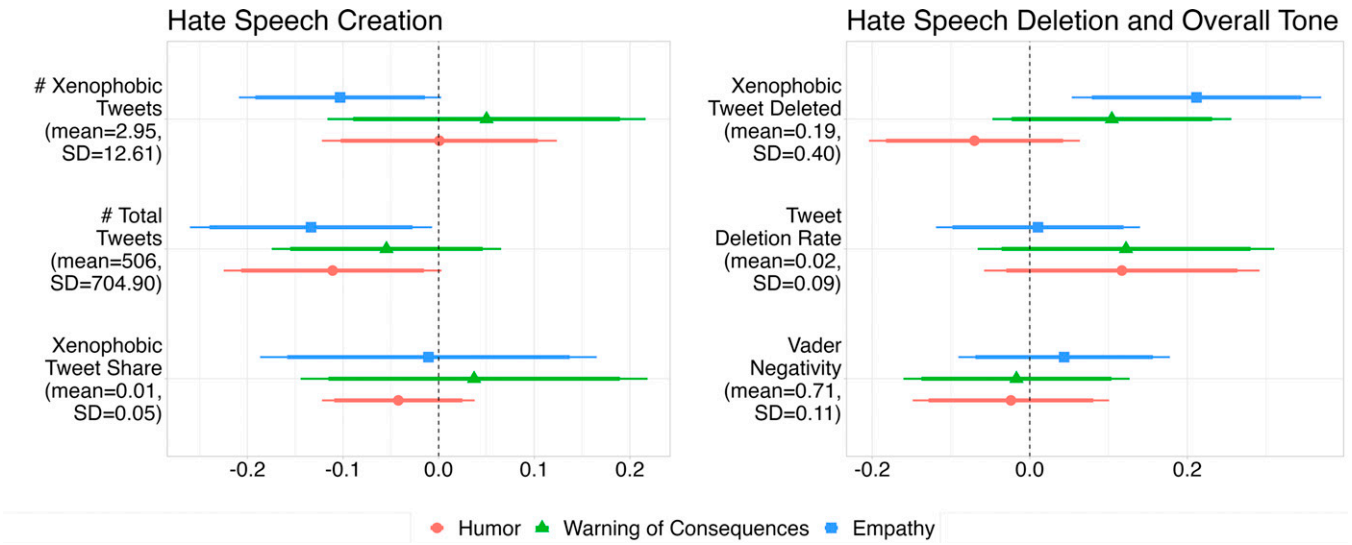

Fig. 2. Point estimates along with 90% and 95% CI from OLS regressions. Effects of the three counterspeech strategies on the "No. of Xenophobic Tweets," "No. of Total Tweets," and the ratio of the two measures ("Xenophobic Tweet Share"), over the 4-wk follow-up period (*Left*); the propensity that the user deleted the original xenophobic tweet ("Xenophobic Tweet Deleted"), the share of retrospectively deleted tweets ("Tweet Deletion Rate") and the average "Vader Negativity" score over the 4-wk follow-up (*Right*). All regression outcomes are standardized to have mean = 0 and SD = 1. Mean and SD of the original, nonstandardized variables are provided in parentheses.

three strategies tested—humor, warning of consequences, and empathy—only the latter yields consistent, albeit relatively small, effects for reducing xenophobic hate speech. This finding points to a central role for empathy in combating online hate speech and echoes previous research that documents how in-person conversations and survey experiments encouraging empathy and perspective taking can reduce hostility toward marginalized groups (13–15). While a focus on empathy seems a promising direction for designing future counterspeech strategies, our results also temper expectations about the effectiveness of some of the most common counterspeech interventions currently deployed by I/NGOs. Future research should consider moving beyond one-shot to repeated interventions, seeking to shed light on the interplay between counterspeaker characteristics and counterspeech messages, and paying attention to unintended "chilling" effects beyond hate speech.

Materials and Methods

This study was approved by the Eidgenössische Technische Hochschule Zurich Ethics Committee (2020-N-155). The experimental design and data collection were preregistered at https://osf.io/km2tc/. Replication materials are available at https://doi.org/10.7910/DVN/ARZ9PU.

Sampling. The field experiment was conducted between November 23, 2020, and January 17, 2021. During the field phase, we collected English-speaking tweets that include one or more words from a dictionary of the most frequently used xenophobic terms and racial slurs (see replication materials). We applied the following preprocessing steps: We excluded all verified user accounts (mostly corporate accounts, organizations, governments, celebrities, or journalists), and excluded retweets (since this study focuses on the creation of hate speech rather than its dissemination), tweets where the targeted minority only appeared in a quoted tweet, tweets that might be perceived as sarcastic, and tweets that were likely sent by bots or minors. We kept a list of Twitter users that entered the study sample, to block reenrollment.

Next, we used dictionary-based sentiment analysis (using the Bing dictionary) to estimate a tweet's ratio of negative to nonnegative words. Working from the most to the least negative, these tweets were then manually classified as containing xenophobic hate speech if their content fulfills at least one of these criteria: 1) uses xenophobic or racist slurs or 2) attacks, denigrates, or strives to discredit an individual or a group that is, or is perceived to be, an immigrant or ethnic/religious minority because of their (perceived) immigrant, minority or outsider status. This sampling yielded 65 to 115 xenophobic tweets per day, providing a total analysis sample of 1,350 Twitter users. Note that the outcomes "Deletion Rate" and "Vader Negativity" are only defined for the 1,149 and 1,279 users who posted at least one tweet preintervention and postintervention, respectively. To maximize sample size, the SUR analysis focuses on the four other outcomes observed for all 1,350 subjects.

Experimental Design. We randomly assigned the sampled Twitter user to one of three treatment arms (with 20% probability each) or the control group (40% probability). We focus on the following three treatments: empathy, where the message is designed to elicit empathy, for example, "For African Americans, it really hurts to see people use language like this"; warning of consequences, where the message is designed to warn the sender about potential social consequences of their tweet, for example, "Hey, remember that your friends and family can see this tweet too"; and humor, with a humorous picture (meme) of an animal, engaging in an obstructing behavior, with captions stating, for example, "It's time to stop tweeting" or "Please stop tweeting."

The treatment intervention consisted of a direct and publicly visible response to the xenophobic tweet and was applied within 24 h from when the tweet was posted. Treatments were administered by one of six human-controlled "bot" accounts, created about 4 wk before the start of the field phase. The accounts did not reveal any information about the user's gender, ethnicity, or nationality; featured a neutral and apolitical posting history; and had about 100 followers. We interdispersed the counterspeech posts with innocuous messages and pictures.

Statistical Analysis. After treatment application, 230 accounts were set to private by the user or suspended by Twitter. These "attrited" accounts are therefore excluded from further analysis. The remaining sample includes n = 1,350 subjects (N_{Humor} = 284, $N_{Warning}$ = 269, $N_{Empathy}$ = 259, and $N_{control}$ = 538). Reassuringly, attrition is balanced across treatment groups: OLS regressions of an attrition indicator on assigned treatment yield very small and statistically not significant differences (humor: $\beta = 0.00$, $P = 0.93$; warning of consequences: $\beta = -0.02$, $P = 0.32$; empathy: $\beta = -0.00$, $P = 0.93$). In addition, 74 tweets (5.5% of the remaining sample) were deleted by the sender after the eligibility checks but before treatment application. To take this noncompliance into account, we use intention-to-treat analysis, implemented with OLS regressions of the outcome of interest on assigned treatment status (rather than the treatment received). The OLS regressions control for pretreatment covariates to increase efficiency and reduce the chance covariate imbalance. We use Lasso-based postdouble selection (12) to select the predictive covariates (and first-order interactions) from the following Twitter account features: language, location, creation date, number of friends and followers, past hate speech and toxicity, and average tweet length. Estimates without covariate adjustment are very similar and provided in the replication materials. The hypotheses motivating the analysis of the outcomes "Xenophobic Tweet Share," "Xenophobic Tweet Deleted," "Tweet Deletion Rate," and "Vader Negativity" were preregistered, while the separate analysis of the two components of "Xenophobic Tweet Share," that is, "No. of Xenophobic Tweets" and "No. of Total Tweets," is exploratory.

Data Availability. Anonymized text data, analysis code, and additional replication materials have been deposited in a Harvard Dataverse at https://doi.org/10.7910/DVN/ARZ9PU (16).

ACKNOWLEDGMENTS. We are grateful to the team at alliance F, whose StopHateSpeech campaign inspired this project, InnoSuisse (Grant 46165.1 IP-SBM), and the Swiss Federal Office of Communications for funding.

1. United Nations, *United Nations Strategy and Plan of Action on Hate Speech – Detailed Guidance on Implementation for United Nations Field Presences* (United Nations, 2020).
2. A. Siegel, *Online Hate Speech in Social Media and Democracy: The State of the Field, Prospects for Reform* (Cambridge University Press, 2020).
3. B. Henson, B. W. Reyns, B. S. Fisher, Fear of crime online? Examining the effect of risk, previous victimization, and exposure on fear of online interpersonal victimization. *J. Contemp. Crim. Justice* **29**, 475–497 (2013).
4. K. Müller, C. Schwarz, Fanning the flames of hate: Social media and hate crime. *J. Eur. Econ. Assoc.* **19**, 2131–2167 (2020).
5. E. Douek, Governing online speech: From 'posts-as-trumps' to proportionality and probability. *Columbia Law Rev.* **121**, 759–834 (2021).
6. I. Gagliardone, D. Gal, T. Alves, G. Martinez, *Countering Online Hate Speech* (United Nations Educational, Scientific and Cultural Organization, 2015).
7. K. Munger, Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Polit. Behav.* **39**, 629–649 (2017).
8. A. A. Siegel, V. Badaan, #No2Sectarianism: Experimental approaches to reducing sectarian hate speech online. *Am. Polit. Sci. Rev.* **114**, 837–855 (2020).
9. K. Munger, Don't @ me: Experimentally reducing partisan incivility on Twitter. *J. Exp. Polit. Sci.* **8**, 102–116 (2020). Erratum in: *J. Exp. Polit. Sci.* **8**, 208 (2020).
10. S. Benesch, D. Ruths, K. P. Dillon, H. M. Saleem, L. Wright, *Considerations for Successful Counterspeech* (Dangerous Speech Project, 2016).
11. C. Hutto, E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text" in *Proceedings of the International AAAI Conference on Web and Social Media*, E. Adar, P. Resnick, Eds. (Association for the Advancement of Artificial Intelligence, 2014), vol. 8, pp. 216–225.
12. A. Belloni, V. Chernozhukov, C. Hansen, Inference on treatment effects after selection among high-dimensional controls. *Rev. Econ. Stud.* **81**, 608–650 (2014).
13. E. L. Paluck, D. P. Green, Prejudice reduction: What works? A review and assessment of research and practice. *Annu. Rev. Psychol.* **60**, 339–367 (2009).
14. D. Broockman, J. Kalla, Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science* **352**, 220–224 (2016).
15. S. Williamson *et al.*, Family matters: How immigrant histories can promote inclusion. *Am. Polit. Sci. Rev.* **115**, 686–693 (2020).
16. G. Gennaro, D. Hangartner, Replication materials for: Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. Harvard Dataverse. https://doi.org/10.7910/DVN/ARZ9PU. Accessed 22 November 2021.