Short of Suspension: How Suspension Warnings Can Reduce Hate Speech on Twitter

Mustafa Mikdat Yildirim, Jonathan Nagler, Richard Bonneau, and Joshua A. Tucker

Debates around the effectiveness of high-profile Twitter account suspensions and similar bans on abusive users across social media platforms abound. Yet we know little about the effectiveness of warning a user about the possibility of suspending their account as opposed to outright suspensions in reducing hate speech. With a pre-registered experiment, we provide causal evidence that a warning message can reduce the use of hateful language on Twitter, at least in the short term. We design our messages based on the literature on deterrence, and test versions that emphasize the legitimacy of the sender, the credibility of the message, and the costliness of being suspended. We find that the act of warning a user of the potential consequences of their behavior can significantly reduce their hateful language for one week. We also find that warning messages that aim to appear legitimate in the eyes of the target user seem to be the most effective. In light of these findings, we consider the policy implications of platforms adopting a more aggressive approach to warning users that their accounts may be suspended as a tool for reducing hateful speech online.

On January 6, 2021, Twitter temporarily suspended Donald Trump's account, requiring him to delete several of his tweets that rejected the election results and appeared to incite violence. Two days later, Twitter permanently suspended his account "due to the risk of further incitement of violence." Trump tried to evade the ban by using the @Potus Twitter account that belongs to sitting U.S. presidents. His attempts were

A list of permanent links to Supplemental Materials provided by the authors precedes the References section.

Data replication sets are available in Harvard Dataverse at: https://doi.org/10.7910/DVN/6FTRZZ

Mustafa Mikdat Yildirim is a fourth year Doctoral Candidate in the Department of Politics at New York University (mmy267@nyu.edu). He studies extremism, hate speech and measures to counter it on social media. He holds bachelor's degrees in political science and economics from Bogazici University, Turkey.

Jonathan Nagler is Professor of Politics and affiliated faculty at the Center of Data Science at New York University, co-director of the Center for Social Media and Politics, and a past president of the Society for Political Methodology, as well as an Inaugural Fellow of the Society for Political Methodology (jonathan.nagler@nyu.edu). Nagler has been a Fernand Braudel Senior Fellow at the European University Institute, a Visiting Scholar at the Russell Sage Foundation, and has taught at Harvard, California Institute of Technology, and the ICPSR and Essex Summer Programs in Political Methodology. He is a co-author of Who Votes Now? *(Princeton University Press, 2014).*

Richard Bonneau is a co-director of the Center for Social Media and Politics, founding member of the Flatiron Institute, and Professor of Biology at New York University (bonneau@nyu.edu). He was selected by Discover *magazine as one of the top twenty scientific minds under forty and a recent review in the top biology journal,* Cell, *lists his 2007 paper on the prediction of global dynamic regulatory networks as a landmark paper in the field of Systems Biology.*

Joshua A. Tucker is Professor of Politics at New York University, Director of New York University's Jordan Center for Advanced Study of Russia, co-Director of the New York University Center for Social Media and Politics (csmapnyu.org), and co-author/editor of the award-winning politics and policy blog The Monkey Cage *at* The Washington Post *(joshua.tucker@nyu.edu). His research focuses on the intersection of social media and politics, as well as mass political behavior in post-communist countries. His most recent book is the co-edited* Social Media and Democracy: The State of the Field *(Cambridge University Press, 2020).*

doi:10.1017/S1537592721002589

unsuccessful as Twitter immediately deleted almost all of his messages. This event is among many that show how the perverse use of social media can increase polarization (Gagliardone et al. 2016, 6-8; Takikawa et al. 2017, 3148) and mobilize inter-group conflict (Bodrunova et al. 2019, 128-129). In order to address these adverse consequences, social media platforms such as Twitter, Facebook, Reddit, and others routinely engage in widespread bans of users (Peters 2020; Guynn 2020; Spangler 2020).

Although account bans are a common measure against hate speech on social media, banning users can have unforeseen consequences such as the migration of banned users to more radical platforms (Livni 2019). After Trump was banned, there were rumors that he himself might start using radical platforms such as Parler or GAB (Guynn 2021), or even start his own platform (Montanaro 2021). Hence, even when bans reduce unwanted deviant behavior within one platform (Chandrasekharan et al. 2017, 14-15), they might fail in reducing the overall deviant behavior within the online sphere.

With this in mind, we draw on political science theory to examine an alternative policy to banning users. More specifically, we test whether warning users of their potential suspension if they continue using hateful language might be able to reduce online hate speech. To do so, we implemented a pre-registered experiment on Twitter in order to test the ability of "warning messages" about the possibility of future suspensions to reduce hateful language online. More specifically, we identify users who are candidates for suspension in the future based on their prior tweets and download their follower lists before the suspension takes place. After a user gets suspended, we randomly assign some of their followers who have also used hateful language to receive a warning that they, too, may be suspended for the same reason.

Since our tweets aim to deter users from using hateful language, we design them relying on the three mechanisms that the literature on deterrence deems as most effective in reducing deviation behavior: costliness, legitimacy, and credibility. In other words, our experiment allows us to manipulate the degree to which users perceive their suspension as costly, legitimate, and credible.

As such, we aim to contribute to a better understanding of countering hate speech on social media. Although there is an increasing prevalence of research that explores the causes, dynamics, consequences, and detection of online hate speech (Müller and Schwarz 2018), we still lack an understanding of the types and effects of interventions aimed at reducing hate-speech. On the one hand, scholars explore the effectiveness of measures that rely on censoring hateful content. On the other hand, there is a burgeoning literature on online speech moderation that, by drawing insights from the study of identity politics, proposes innovative interventions that reduce people's likelihood of spreading hate speech (Munger 2017, 2020; Siegel and Badaan 2020).

By testing the relative effectiveness of suspension warnings designed to highlight costliness, credibility, and legitimacy, we contribute to a better understanding of the exact mechanisms of deterrence that are most effective in reducing deviant behaviors online. To our knowledge, these mechanisms have not previously been analyzed in a naturalistic setting with real-time tracking of subject behavior.

This study also contributes to works that explore the impact of user bans by online social platforms. Although some studies show that bans are effective in reducing the overall levels of hate speech in one platform (Chandrasekharan et al. 2017, 14-15), they rarely follow users' subsequent behavior and are, as such, not informative of the overall impact that bans can have on shaping the behavior of users who remain on the platform.

Our study provides causal evidence that the act of sending a warning message to a user can significantly decrease their use of hateful language as measured by their ratio of hateful tweets over their total number of tweets. Although we do not find strong evidence that distinguishes between warnings that are high versus low in legitimacy, credibility, or costliness, the high legitimacy messages seem to be the most effective of all the messages tested. We also test for a set of heterogeneity effects—number of followers, anonymity of the profile, level of Twitter engagement—and do not find evidence that the results are driven by any of these profile characteristics.[1]

This reflection is organized as follows. First, we discuss the relevant theoretical underpinnings of our hypotheses and the literature on deterrence. Next, we present the details of our innovative experimental design, followed by our results. We then consider the policy implications of platforms adopting a more aggressive approach to warning users that their accounts may be suspended as a tool for reducing hateful speech online.

Deterring Hate Speech: Credibility, Costliness, and Legitimacy

There is a large body of literature that studies which types of interventions are most effective in reducing prejudice and conflict in real-world settings (Paluck and Green 2009a; Broockman and Kalla 2016; Munger 2017; Siegel and Badaan 2020). However, these works largely focus on inter-group dynamics, and draw on social psychological theories related to the salience of group identity.

In our study, we isolate components of interventions that are not related to identity dynamics. Instead, we explore factors that make a warning message effective in reducing hateful behavior. To begin, there is a large and still growing body of work from the literature on deterrence that provides evidence that sending warning

https://doi.org/10.1017/S1537592721002589 Published online by Cambridge University Press

messages in cyberspace represents an important avenue for deterring individuals from malevolent behavior (Wilson et al. 2015; Silic et al. 2016; Testa et al. 2017).

Scholars argue that warning messages of punishment can take on two different forms of deterrence: general and specific. General deterrence is based on one's vicarious experiences with punishment and punishment avoidance, whereas specific deterrence refers to the effect of one's personal experiences with punishment and punishment avoidance (Stafford and Warr 1993, 127). In our study, our warning messages are meant to have a general deterrence effect because we make hateful users aware of the punishment other users were exposed to due to their use of hateful language.

When it comes to the deterrent effects of our warning messages, we are interested in the degree to which we are able to reduce users' hateful language. Deterrence scholars distinguish between the ability of sanction threats to eradicate completely the deviant behavior (absolute deterrence), versus the effect of sanctions in reducing the severity and frequency of individual offending (restrictive deterrence) (Gibbs 1968, 518; Jacobs 2010, 423). Our experiment aims to test restrictive deterrence as we do not think that sending a single warning tweet to a hateful user would stop their use of hateful language completely.

Multiple factors are needed to make a warning message of punishment restrictively deter its targets from engaging in deviant behavior. As a first step, the deterring message should be conveyed to the target audience for it to be deterrent (Geerken and Gove 1974, 499), or based on Communication-Human Information Processing (C-HIP) Model (Conzola and Wogalter 2001, 312; Wogalter 2006, 34-39), the warning should be communicated from the source (person or entity delivering the message) to the receiver. The delivery itself is not enough. It should also get the receiver's attention. Once the message gets the receiver's attention, the receiver should understand what the warning message says. Next, the warning message should change the receiver's attitudes and beliefs about the costs and benefits of their deviant behavior (Beccaria 1963, 59, 94; Paternoster 1987, 174-175). Finally, the individual must understand what actions can be taken to avoid the costs of their unwanted behavior (Rogers 1975, 97-98) to change their behavior as a result of receiving a warning message.

In our study, we effectively deliver our warning message to hateful users by sending public tweets to their profiles.[2] In our context, punishment is account suspension. Since our tweets are different from a usual tweet that a user would receive from other users, and since the user gets notifications when they receive a tweet from another user, we presume that our tweets get their attention. We also avoid any type of jargon within the language of our warning tweets to avoid the risk that a user would misunderstand or not understand our tweets. We conduct manipulation checks to make sure that our tweets convey what we want them to convey.[3]

We clearly express in our warning tweets the potential adverse consequences of the target users' behavior—suspension from Twitter—and make them aware that people they followed faced these consequences. Having established the conditions that would make a warning message deterrent based on the literature, and based on the corroborating evidence from recent works that study the effect of surveillance warning banners on the behavior of trespassers (Stockman, Heile, and Rein 2015; Wilson et al. 2015), we pre-registered the following hypothesis:

> H1. A tweet that warns a user of a potential suspension in the case of employing hate speech will lead that user to decrease their use of hateful language.

We design our messages based on the literature of deterrence. Theories on deterrence suggest three main channels. The first is the costliness, which emphasizes the influence of the perceptions of sanctions' severity in generating effective deterrence (Cusson 1993; Gibbs 1968; Paternoster 1987). The second is credibility (Nagin 1998, 8), where the user's conviction regarding the probability that a threat will occur affects how much they will be deterred from their unwanted behavior (Rogers 1975, 97). One factor that can make warnings credible is the authority of the source. Kiesler et al. (2012, 133-134) point out that moderation attempts on online platforms by members who seem to deserve to be in the moderator position are considered as being more credible by other members. The third channel is the legitimacy[4] of the warning message. Sherman (1993, 445) argues that the legitimacy of experienced punishment is essential for the acknowledgement of shame, which then conditions deterrence.

Based on these three different channels, we pre-registered three pairs of warning tweets. Each pair emphasizes high and low versions of each channel. For example, among the pair of warning tweets that we designed based on costliness, low-cost tweets trigger the costliness of the unwanted behavior less than the high-cost tweets do.[5]

Deterrence as a Function of the Target's Features

There are also reasons why we might expect to find differential effects based on user characteristics. We expect a greater cost from suspension for users who are more heavily invested in their profile (as measured by the number of tweets they post, the number of followers they have, or the age of their profile). Also, users who are anonymous (i.e., users who do not reveal their names or photos in their profile) would be expected to be less sensitive to our warning messages because anonymous users' perceived risk of detection would be lower (Munger 2017, 630-631).[6]

https://doi.org/10.1017/S1537592721002589 Published online by Cambridge University Press

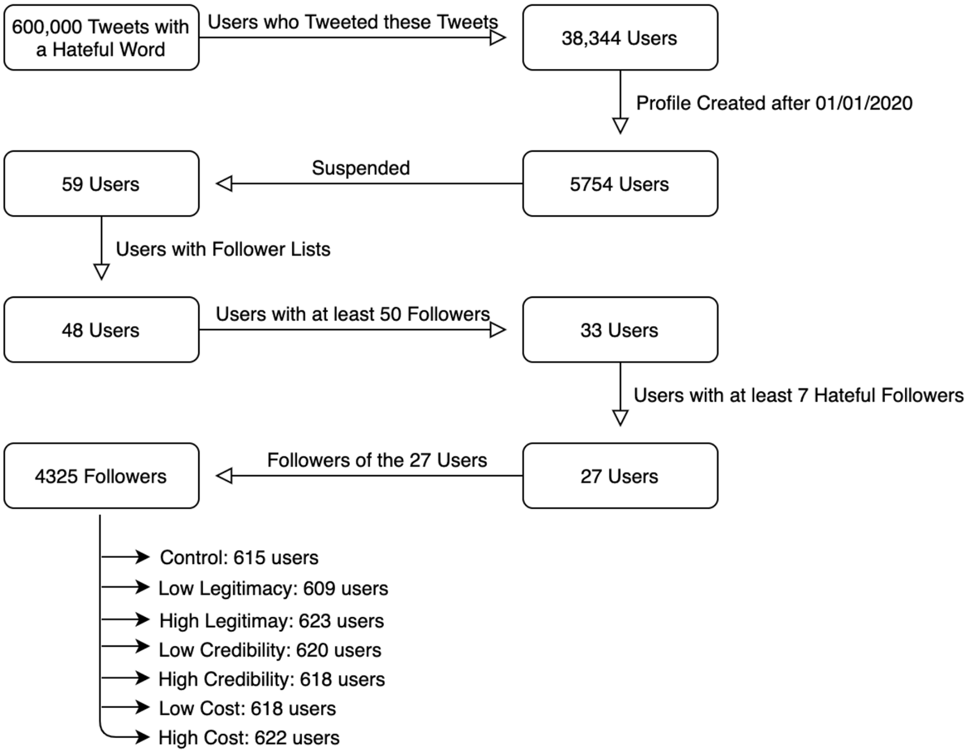
Experimental Design

As we argue in the previous section, to effectively convey a warning message to its target, the message needs to make targets aware of the consequences of their behavior, and also make them believe that these consequences will be administered (Geerken and Gove 1974). Therefore, we designed experiments that would a) make Twitter users aware of the fact that their account could possibly be suspended, but at the same time, b) only sent these warnings to people who could credibly believe that their account could be suspended. To ensure that this second condition held, we limited our participant population to people who had previously used hateful language on Twitter *and* followed someone who actually had just been suspended.[7] In order to measure the effectiveness of our interventions, we could then compare the use of hateful speech by those who had received a warning with those who had not.

More specifically, we used a pre-registered design, the broad contours of which are illustrated in figure 1.[8] Our first step was to find accounts that could possibly be suspended during the term of our study. To identify such "suspension candidates", we began by downloading 600,000 tweets on July 21, 2020 that were posted in the week prior and that contained at least one word from the hateful language dictionary created by Munger (2017).[9] During the period, Twitter was flooded by hateful tweets against the Asian and Black communities due to Covid and BLM protests, respectively (Kumar and Pranesh 2021; Ziems et al. 2020).

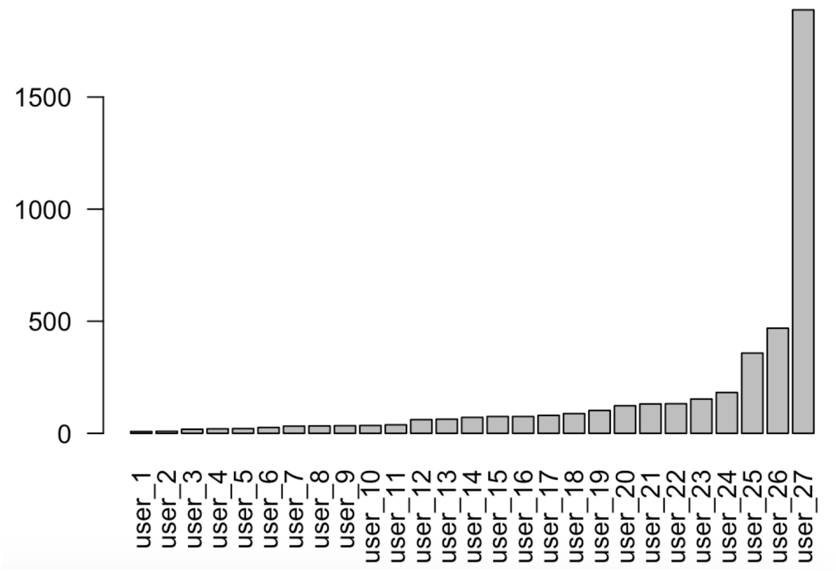
Next, we downloaded the IDs of the 38,444 users who tweeted these tweets. We filtered these users to the 5,754 users who created their profile after January 1, 2020. Our reasoning here was that newly created users would be more likely to be suspended as compared to people with older accounts. We next downloaded all the followers of the 5754 users because we were expecting some of these users to get suspended and wanted to obtain the list of their followers before they got suspended. Over the course of fourteen days, 59 out of these 5,754 users did in fact get suspended. Out of the 59 users, we were able to download the follower lists of 48 of them before they were suspended. Out of those 48 users, we included only those with more than 50 followers to be able to randomize their followers into six treatment groups and a control group, which decreased the number of "seed users" from 48 users to 33 users with 39,659 followers. We downloaded the most recent 800 tweets of each of these 39,659 followers, and then

Figure 1
Research design


600,000 Tweets with a Hateful Word
Users who Tweeted these Tweets
38,344 Users
Profile Created after 01/01/2020
5754 Users
Suspended
59 Users
Users with Follower Lists
48 Users
Users with at least 50 Followers
33 Users
Users with at least 7 Hateful Followers
27 Users
Followers of the 27 Users
4325 Followers
Control: 615 users
Low Legitimacy: 609 users
High Legitimay: 623 users
Low Credibility: 620 users
High Credibility: 618 users
Low Cost: 618 users
High Cost: 622 users

https://doi.org/10.1017/S1537592721002589 Published online by Cambridge University Press

Figure 2
Number of hateful followers per suspended user



Note: Names of the suspended users are anonymized to preserve privacy.

calculated the percentage of each follower's tweets that used at least one hateful term from Munger's (2017) dictionary. We then filtered these followers to select only those who used a hateful term in at least 3% of their tweets over a month from July 4 to August 4. Munger (2017, 635) shows that among randomly sampled users, those who are at the seventy-fifth and higher percentile in terms of using hateful language have at least 3 percent of their tweets with hateful language, and calls the level of 3% the "regularly offensive threshold." We label these followers as "hateful followers." We filtered our 33 seed users to users with at least 7 "hateful" followers so that we could have at least one follower in each treatment condition from each seed user. This resulted in 27 suspended users with a total of 4,327 followers, who were then randomly assigned to one of our six treatment groups and to our control.[10]

Figure 2 shows the distribution of the followers in all seven conditions (six treatment arms and one control arm to which we did not send a tweet). Most suspended users had somewhere between 100–500 hateful followers whose tweets rise above the "regularly offensive threshold," although one suspended user had many more hateful followers compared to the others (1,889 followers).

Table 1 shows the summary statistics of the suspended users' 4,327 followers who tweet hateful language in at least 3% of their tweets. The mean proportion of hateful tweets is 6%, which is twice the ratio that Munger (2017) labels as a regular offensiveness threshold. Although the mean number of followers is very high, the median is much lower, reflecting the fact that the distribution is highly skewed towards accounts with lower numbers of followers. Activity is the daily number of tweets, which shows that the average user in our sample tweets eight times per day. Anonymity score is a variable that takes values of 0, 1, or 2. If a user's anonymity score is equal to 0, the user has their own photo in their profile and their own name as their username. If the anonymity score is 1, they have either of the two. If it is 2, the user has neither and is considered completely anonymous.

After randomizing the followers into six treatment groups and a control group, we sent one of six tweets (representing the six theoretically informed treatment groups) from six separate accounts that we created.[11] We did not send any tweets to the control group.[12] The six tweets that we designed are meant to manipulate the costliness of the suspension in the eyes of the treated, the extent to which they perceive our warning as legitimate, and the degree to which they perceive our warning as credible. These messages can be seen in figure 3.

Results

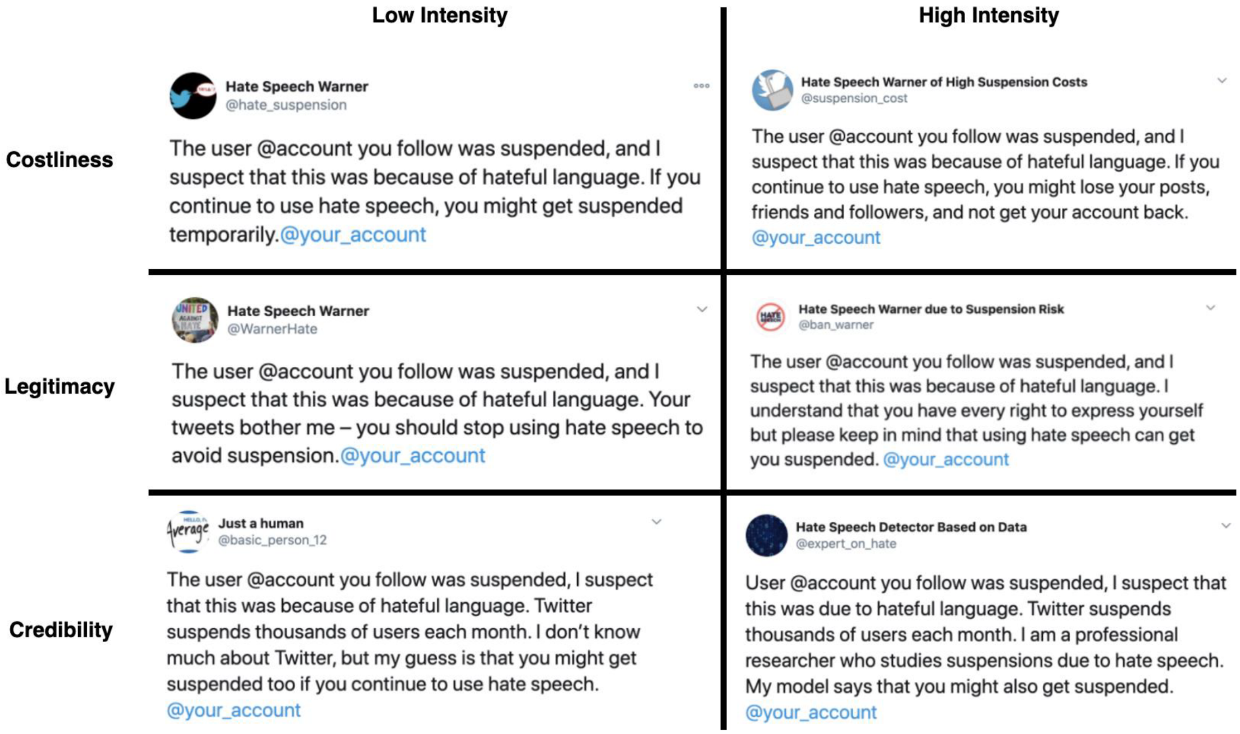
We next present the key results of our experimental analyses; additional analyses described in our preregistration plan can be found in the online appendix.[13] The coefficient plot in figure 4 shows the effect of sending any type of warning tweet on the ratio of tweets with hateful language over the tweets that a user tweets. The outcome variable is the ratio of hateful tweets over the total number of tweets that a user posted over the week and

https://doi.org/10.1017/S1537592721002589 Published online by Cambridge University Press

Table 1
Summary statistics on followers of the suspended users

	Number of Followers	Activity	Age	Ratio of Hateful Tweets	Anonymity
Mean	2898	8.3	6.2	0.07	1.2
Median	854	5.5	6.4	0.05	1
SD	8860	7.7	1.5	0.08	0.7
Min	0	0.03	2.6	0.03	0
Max	268891	25.67	8.5	1	2

Figure 3
Treatment tweets



Hate Speech Warner
@hate_suspension

The user @account you follow was suspended, and I suspect that this was because of hateful language. If you continue to use hate speech, you might get suspended temporarily.@your_account

Hate Speech Warner of High Suspension Costs
@suspension_cost

The user @account you follow was suspended, and I suspect that this was because of hateful language. If you continue to use hate speech, you might lose your posts, friends and followers, and not get your account back. @your_account

Hate Speech Warner
@WarnerHate

The user @account you follow was suspended, and I suspect that this was because of hateful language. Your tweets bother me – you should stop using hate speech to avoid suspension.@your_account

Hate Speech Warner due to Suspension Risk
@ban_warner

The user @account you follow was suspended, and I suspect that this was because of hateful language. I understand that you have every right to express yourself but please keep in mind that using hate speech can get you suspended. @your_account

Just a human
@basic_person_12

The user @account you follow was suspended, I suspect that this was because of hateful language. Twitter suspends thousands of users each month. I don't know much about Twitter, but my guess is that you might get suspended too if you continue to use hate speech. @your_account

Hate Speech Detector Based on Data
@expert_on_hate

User @account you follow was suspended, I suspect that this was due to hateful language. Twitter suspends thousands of users each month. I am a professional researcher who studies suspensions due to hate speech. My model says that you might also get suspended. @your_account

Note: As a manipulation check, we ran an MTurk experiment with a separate sample of fifty people to Refer to online appendix E for a more detailed discussion.

month following the treatment. The effects thus show the change in this ratio as a result of the treatment.

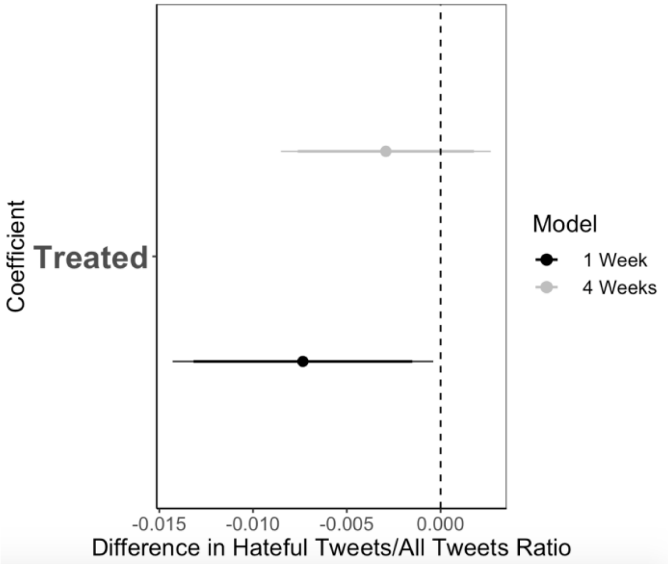
We find support for our first hypothesis: a tweet that warns a user of a potential suspension will lead that user to decrease their ratio of hateful tweets by 0.007 for a week after the treatment. Considering the fact that the average pre-treatment hateful tweet ratio is 0.07 in our sample, this means that a single warning tweet from a user with 100 followers reduced the use of hateful language by 10%. We suspect as well that these are conservative estimates, in the sense that increasing the number of followers that our account had could lead to even higher effects, as Munger (2017) and Siegel and Badaan (2020) show in their studies, to say nothing of what an official warning from Twitter would do.[14]

The coefficient plot in figure 5 shows the effect of each treatment on the ratio of tweets with hateful language over the tweets that a user tweets. Although the differences across types are minor and thus caveats are warranted, the most effective treatment seems to be the high legitimacy tweet; the legitimacy category also has by far the largest difference between the high- and low-level versions of the three categories of treatment we assessed. Interestingly, the tweets emphasizing the cost of being suspended appear to be the least effective of the three categories; although the effects are in the correctly predicted direction, neither of the cost treatments alone are statistically distinguishable from null effects.

An alternative mechanism that could explain the similarity of effects across treatments—as well as the costliness

https://doi.org/10.1017/S1537592721002589 Published online by Cambridge University Press

Figure 4
The effect of sending a warning tweet on reducing hateful language



Note: See table G1 in online appendix G for more details on sample size and control coefficients.

channel apparently being the least effective—is that perhaps instead of deterring people, the warnings might have made them more reflective and attentive about their language use. Such a mechanism would be consistent with prior disinformation studies that demonstrated that nudging, flagging, or alerting users to the possibility of inaccuracy makes users more attentive to questions of accuracy (Pennycook et al. 2021). If that is the case, then perhaps our act of warning people impacted their behavior simply by causing them to be more reflective about their own actions, as opposed to motivating a change in behavior out of fear of possible punishment.

Discussion and Implications

Our results show that only one warning tweet sent by an account with no more than 100 followers can decrease the ratio of tweets with hateful language by up to 10%, with some types of tweets (high legitimacy, emphasizing the legitimacy of the account sending the tweet) suggesting decreases of perhaps as high as 15%–20% in the week following treatment. Considering that we sent our tweets from accounts that have no more than 100 followers, the effects that we report here are conservative estimates, and could be more effective when sent from more popular accounts (Munger 2017).

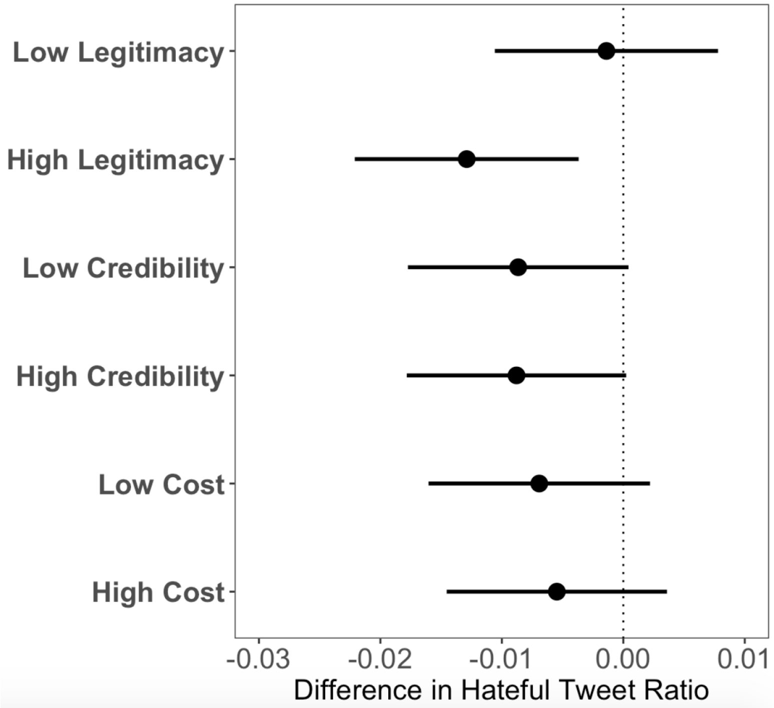
In reducing hateful language, our paper builds on works from political science that explore various interventions that reduce intergroup prejudice and conflict (Paluck and Green 2009b; Samii 2013; Simonovits, Kezdi, and Kardos 2018; Kalla and Broockman 2020). These strategies mostly rely on intergroup contact theory (Pettigrew 1998), on interpersonal conversations (Kalla and Broockman 2020), or on making subjects play a perspective taking game where they practice thinking from the perspective of the outgroup members (Simonovits, Kezdi, and Kardos 2018) in order to decrease antipathy towards other groups.

A recently burgeoning literature shows that online interventions can also decrease behaviors that could harm the other groups by tracking subjects' behavior over social media. These works rely on online messages on Twitter that sanction the harmful behavior, and succeed in reducing hateful language (Munger 2017; Siegel and Badaan 2020), and mostly draw on identity politics when designing their sanctioning messages (Charnysh et al. 2015). We contribute to this recent line of research by showing that warning messages that are designed based on the literature of deterrence can lead to a meaningful decrease in the use of hateful language without leveraging identity dynamics. This is important because interventions that rely on identity dynamics require the knowledge of the target user's identity to be effective (Munger 2017; Siegel and Badaan 2020). However, obtaining such knowledge is not always feasible, as it is not uncommon for people on social media platforms such as Twitter to have anonymous profiles. Even when profiles are non-anonymous, labeling them to design an effective intervention that is based on the identity of the target user may be costly and time consuming.

Our findings suggest, therefore, that one option for reducing hate speech on Twitter would be to warn users who have reason to suspect that they might be at risk of

https://doi.org/10.1017/S1537592721002589 Published online by Cambridge University Press

Figure 5
Reduction in hate speech by treatment type



Note: See table G2 in online appendix G for more details on sample size and control coefficients.

being suspended by Twitter for using hateful language. In our experiment, this "at risk" category was based on the three-fold combination of finding users who themselves use hateful language *and* followed someone who got suspended from the platform for using hateful language *and* alerting them to the fact that the person they followed (known in Twitter parlance as a "friend") had been suspended.

How might this be done? Two options are worthy of discussion: relying on civil society or relying on Twitter. Our experiment was designed to mimic the former option, with our warnings mimicking non-Twitter employees acting on their own with the goal of reducing hate speech/protecting users from being suspended. From our intervention, it seems that at a bare minimum, such warnings can result in a short-term reduction in hate speech on Twitter, which would seem to be normatively desirable. And while we did not find longer-term effects from a single warning, it is possible that different variations of this stimulus (e.g., multiple warnings over an extended time period) could have a longer-term effect. But even if not, the cumulative effect of short-term reductions in hate speech—if new warnings were issued to new people with regularity—would still reduce hate speech on the platform.

The question, of course, is how such a program could be implemented at the scale of Twitter. It did take a non-trivial amount of work and technical skill for us to design and implement our interventions. While it is certainly possible that an NGO or a similar entity could try to implement such a program, the more obvious solution would be to have Twitter itself implement the warnings. After all, Twitter has access to all of the necessary data: the company knows exactly who has been suspended and when, who their followers are, and whether or not those users have crossed the "regularly offensive" threshold. Moreover, Twitter has the capacity to completely automate this process, which means that it can be applied at scale; it also means that Twitter could easily run much more extensive versions of our study to hone in on the most effective types of warnings.

Indeed, Twitter has also recently shared publicly results from its own testing of a different form of warning. More specifically, the company reported "testing prompts in 2020 that encouraged people to pause and reconsider a potentially harmful or offensive reply—such as insults, strong language, or hateful remarks—before Tweeting it. Once prompted, people had an opportunity to take a moment and make edits, delete, or send the reply as is."[15] This appears to result in 34% of those prompted electing

https://doi.org/10.1017/S1537592721002589 Published online by Cambridge University Press

either to review the Tweet before sending, or not to send the Tweet at all.

We note three differences from this endeavor. First, in our warnings, we try to reduce people's hateful language *after* they employ hateful language, which is not the same thing as warning people *before* they employ hateful language. This is a noteworthy difference, which can be a topic for future research in terms of whether the dynamics of retrospective versus prospective warnings significantly differ from each other. Second, Twitter does not inform their users of the examples of suspensions that took place among the people that these users used to follow. Finally, we are making our data publicly available for re-analysis.

We stop short, however, of unambiguously recommending that Twitter simply implement the system we tested without further study because of two important caveats. First, one interesting feature of our findings is that across all of our tests (one week versus four weeks, different versions of the warning—figures 2 (in text) and A1(in the online appendix)) we never once get a positive effect for hate speech usage in the treatment group, let alone a statistically significant positive coefficient, which would have suggested a potential *backlash* effect whereby the warnings led people to become more hateful. We are reassured by this finding but do think it is an open question whether a warning from Twitter—a large powerful corporation and the owner of the platform—might provoke a different reaction. We obviously could not test for this possibility on our own, and thus we would urge Twitter to conduct its own testing to confirm that our finding about the lack of a backlash continues to hold when the message comes from the platform itself.[16]

The second caveat concerns the possibility of Twitter making mistakes when implementing its suspension policies. Now, of course, these policies already exist and are being implemented, so mistakes in the process already cause harm to users whose accounts are incorrectly suspended. However, implementing the warning system we tested in our experiment would in a sense be broadcasting the fact of that suspension—and attributing a reason for it —to a larger number of users. We were careful in our experiments to say that we *suspected* the account was suspended because of hateful language (which was absolutely true—we did suspect this but did not know definitively), but coming from Twitter such ambiguity would likely be less credible.[17] Thus it would be important to weigh the incremental harm that such a warning program could bring to an incorrectly suspended user (importantly, beyond the harm that the already existing suspension policy is causing) versus the benefit of the incremental decrease in hate speech on the platform. We suspect the dispersed benefits would outweigh the concentrated harm, but in order to definitively feel comfortable with this conclusion we would want to see Twitter's data about how often accounts that are suspended for hate speech are found to have been incorrectly suspended, as well as whether there are disproportionate numbers of incorrect suspensions across different socio-demographic groups within society. Nevertheless, it is worth considering whether it would be better for Twitter—should it decide to test/implement a version of what we have done—to anonymize the suspended user in the warning tweet (e.g., "someone you follow was suspended" as opposed to the "@[user] was suspended" we employed). This might help mitigate the potential harm, but such an approach would clearly need to be tested to see if it still has the same impact on reducing hateful speech.

While our experiment was conducted solely on Twitter, there is nothing inherent about the idea of using warnings of suspended friends to try to reduce hateful speech that limits such an approach to Twitter. However, it is worth highlighting that there are particulate affordances of Twitter that make the platform amenable to this sort of intervention, namely the fact that the users are enmeshed in networks and that activity on the platform is largely public. The former raises our expectation that people will care that user X was suspended because the user already has a previous relationship with user X (i.e., the user chose to follow user X), so it might be the case that simply learning some random user was suspended (on a platform such as Reddit where there are not follower relationships) might not be as effective. The fact that Twitter is a public platform may also have made users less surprised to see a warning from a random account such as ours, which might not be the case on a platform where users are more accustomed to thinking their posts are private, such as Facebook, although this concern might be less of an issue if the message is coming from the platform itself.[18]

Despite these caveats, our findings suggest that hate-speech moderations can be effective without priming the salience of the target users' identity. Explicitly testing the effectiveness of identity versus non-identity motivated interventions will be an important subject for future research.

Acknowledgements

Yildrim and Tucker designed the research; Yildrim performed the research; Yildrim, Nagler, Bonneau, and Tucker planned the analyses; Yildrim analyzed data and wrote the first draft of the paper, and all authors contributed to revisions. The authors are thankful to the New York University Center for Social Media and Politics (CSMaP) weekly meetings and the New York University Comparative Politics Workshop for their helpful feedback. The Center for Social Media and Politics for New York University is generously supported by funding from the National Science Foundation, the John S. and James L. Knight Foundation, the Charles Koch Foundation, the Hewlett Foundation, Craig Newmark Philanthropies, the Siegel Family Endowment, and New York University's Office of the Provost.

https://doi.org/10.1017/S1537592721002589 Published online by Cambridge University Press

Supplementary Materials

To view supplementary material for this article, please visit http://doi.org/10.1017/S1537592721002589.

Appendix A. Missing Observations

Appendix B. Power Analysis

Appendix C. Heterogeneous Treatment Effects

Appendix D. Heterogeneous Treatment Effects with Continuous Variables

Appendix E. Manipulation Checks for Our Warning Tweets

Appendix F. Alternative Outcome Variable

Appendix G. Detailed Tables for the Figures in the Paper

Appendix H. Treatment Tweets and Profiles that Sent Them

Appendix I. Pre-Registered Hypotheses and Their Theoretical Justifications

Appendix J. Dictionary of Hateful Words

Notes

1 In our pre-analysis plans, we specified hypotheses about the relative impact of high versus low levels of costliness, legitimacy, and credibility. Due to the null findings regarding these differential effects within each type of warning and at the request of the editors, we have relegated the discussion of these hypotheses to the online appendix I and include the results of these analyses in online appendices C and D.

2 On Twitter, it is possible to send a private message to a user only if the user permits private messages from accounts that they do not follow. As such, we elected to send all of our messages publicly as tweets. Had it been possible to send private messages, we would likely have run a treatment arm(s) using private messages and would have theorized in our pre-analysis plan about expected differences across public and private messages. As this was not possible, we did not address the topic in the pre-analysis plan, and consequently only address the nature of public messages in our discussion section when we reference the potential harms from using public messages.

3 We ran an MTurk experiment with a separate sample of fifty people to ensure that our tweets convey what we want them to convey. Refer to online appendix E for a more detailed discussion.

4 In retrospect, we have realized "problematic" or "offensive" is an alternative interpretation of how we operationalized the extent to which our tweet is "legitimate." However, we kept the label "legitimacy" because we report the mechanism as "legitimacy" in our Pre-Analysis Plan.

5 The theoretical justification for each channel, and the three additional hypotheses that we draw can be found in online appendix I. In our Pre-Analysis Plan, we specified that increasing costliness, legitimacy, or credibility of a warning tweet would lead to higher deterrence than the corresponding lower-level version of that channel. All of these hypotheses returned null results. In the interest of space and at the advice of the editors, we do not report the results of these tests in the body of the paper. However, in the interest of transparency and avoiding the file drawer problem (Franco, Malhotra, and Simonovits 2014), as well as in order to make sure our findings are included in any future meta-analyses, we include our results in online appendix G, where we report the null findings.

6 The hypotheses on differential effects were a part of our pre-analysis plan. All of these hypotheses returned null results. In the interest of space and at the advice of the editors, we do not report the results of these tests in the body of the paper. However, in the interest of transparency and avoiding the file drawer problem (Franco, Malhotra, and Simonovits 2014), as well as in order to make sure our findings are included in any future meta-analyses, we include our results in online appendices C and D, where we report the null findings.

7 Twitter outlines three specific categories—1) safety, (2) privacy, (3) authenticity, each of which entails finer subcategories on specific violating activities (see the link for at the end of this paragraph). We specifically focus on the sub-categories of "hateful conduct" and "abuse/harassment" under the category "safety." "Hateful conduct" is defined as "promoting violence against, threatening, or harassing other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease." "Abuse/harassment" is defined as "targeted harassment of someone, or incitement of other people to do so." When a user Tweets in violation of these categories, Twitter can temporarily or permanently suspend their account. The user can apply to Twitter in order to get their account back. Although Twitter does not release systematic reports on the rate of the suspensions, a recent study finds that the rate of suspensions is around 1.5 percent (Chowdhury et al. 2020), which is similar to our study where approximately 1% of the accounts that we tracked were suspended. The link for the policy is (https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy).

8 Please see our Pre-Analysis Plan: https://osf.io/jtpq3/?view_only=5d99dc5eab90432e97ad132a5837bbd7.

9 We use Munger's 2017 dictionary, which highlights race dynamics. This allowed us to capture a high number of hateful tweets during a period where hateful language against Black Lives Matter protests was prevalent (Kumar and Pranesh 2021). We also want to note that to the extent that we identify false

https://doi.org/10.1017/S1537592721002589 Published online by Cambridge University Press

positives, that should bias us against finding any effect for our treatment, which gives us greater confidence in our positive findings and some additional reason for caution in our null findings. For more detail, see the dictionary in online appendix J.

[10] The results of the power analysis that we conducted so that we could decide on the lowest number of followers to collect in order to be sufficiently powered to detect hypothesized effects can be found in online appendix B.

[11] We created profiles that had around 100 followers for this experiment. Considering that previous studies have used profiles with much higher number of followers (e.g., >500; see Munger 2017 and Badaan and Siegel 2020), the effects that we find can be argued to be conservative estimates, as accounts with higher number of followers would be perceived as being more popular, therefore expected to be more influential; Munger 2017.

[12] Except for the low legitimacy tweet where we say that we do not know much about Twitter, none of our tweets contain any deceptive elements. We also think that advising people to reduce their hateful language with a warning tweet does not violate the norms of ethical research. We do not see any harm from using such measures beyond what users might encounter in the course of their own activity on Twitter; the experiment was approved by the New York University Institutional Review Board #IRB-FY2020-4465.

[13] Refer to online appendix G.

[14] As we discuss later, though, we cannot rule out the possibility that an official Twitter warning might inspire backlash that our "concerned user" tweets have not.

[15] https://blog.twitter.com/en_us/topics/product/2021/tweeting-with-consideration. Retrieved on July 15, 2021.

[16] Should Twitter choose to conduct such studies, we would suggest they do so in a transparent manner, preregistering the research design beforehand and agreeing to make the results available to researchers for replication analyses.

[17] It is possible, though, that Twitter could find some more ambiguous language as to the reason for the suspension, which might leave open the possibility that the account was suspended for hate speech violations without saying it explicitly, such as by noting "An account you follow has recently been suspended. We suspect that you, too, are at risk of being suspended for using hateful language" or something to this effect.

[18] Logistically, we could not even run such an experiment on our own on Facebook, where we would not have access to people's private accounts.

References

Beccaria, Cesare. 1963 [1764]. "On Crimes and Punishments." Trans. H. Paolucci. Indianapolis, IN: Bobbs-Merrill.

Bodrunova, Svetlana S., Ivan Blekanov, Anna Smoliarova, and Anna Litvinenko. 2019. "Beyond Left and Right: Real-World Political Polarization in Twitter Discussions on Inter-Ethnic Conflicts." *Media and Communication* 7(3): 119–32. https://doi.org/10.17645/mac.v7i3.1934

Broockman, David, and Joshua Kalla. 2016. "Durably Reducing Transphobia: A Field Experiment on Door-to-Door Canvassing." *Science* 352(6282): 220–24.

Chandrasekharan, Eshwar, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. "You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined through Hate Speech." *Proceedings of the ACM on Human-Computer Interaction* 1(CSCW): 1–22.

Charnysh, Volha, Christopher Lucas, and Prerna Singh. 2015. "The Ties that Bind: National Identity Salience and Pro-Social Behavior toward the Ethnic Other." *Comparative Political Studies* 48(3): 267–300.

Chowdhury, Farhan Asif, Lawrence Allen, Mohammad Yousuf, and Abdullah Mueen. 2020. "On Twitter Purge: A Retrospective Analysis of Suspended Users." In *Companion Proceedings of the Web Conference 2020*, 371-378. https://doi.org/10.1145/3366424.3383298

Conzola, Vincent C., and Michael S. Wogalter. 2001. "A Communication–Human Information Processing (C–HIP) Approach to Warning Effectiveness in the Workplace." *Journal of Risk Research* 4(4): 309–22.

Cusson, Maurice. 1993. "Situational Deterrence: Fear during the Criminal Event." *Crime Prevention Studies* 1(3): 55–68.

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. "Publication Bias in the Social Sciences: Unlocking the File Drawer." *Science* 345(6203): 1502–505.

Gagliardone, Iginio, Matti Pohjonen, Zenebe Beyene, Abdissa Zerai, Gerawork Aynekulu, Mesfin Bekalu, Jonathan Bright *et al.* 2016. "Mechachal: Online Debates and Elections in Ethiopia—From Hate Speech to Engagement in Social Media." *Available at SSRN 2831369* (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2831369).

Geerken, Michael R., and Walter R. Gove. 1974. "Deterrence: Some Theoretical Considerations." *Law and Society Review* 9(3): 497–513.

Gibbs, Jack P. 1968. "Crime, Punishment, and Deterrence." *Southwestern Social Science Quarterly* 48(4): 515–30.

https://doi.org/10.1017/S1537592721002589 Published online by Cambridge University Press

Guynn, Jessica. 2020. “Facebook Ranks Deleting Anti-Black and ‘Most Harmful’ Hate Speech over Comments about White People and Men.” *USA Today*, December 3. Retrieved March 1, 2021 (https://www.usatoday.com/story/tech/2020/12/03/facebook-ranks-hate-speech-black-over-attacks-white-people-men/3813931001/).

——. 2021.“Donald Trump Ruled Facebook, Twitter before He Was Banned. Will @realdonaldtrump Log into Gab or Somewhere Else?” *USA Today*, February 8. Retrieved March 1, 2021 (https://www.usatoday.com/story/tech/2021/02/08/trump-facebook-twitter-youtube-ban-where-next-gab-parler/4440645001/).

Jacobs, Bruce A. 2010. “Deterrence and Deterrability.” *Criminology* 48(2): 417–41.

Kalla, Joshua L., and David E. Broockman. 2020. “Reducing Exclusionary Attitudes through Interpersonal Conversation: Evidence from Three Field Experiments.” *American Political Science Review* 114(2): 410–25.

Kiesler, Sara, Robert Kraut, Paul Resnick, and Aniket Kittur. 2012. “Regulating Behavior in Online Communities.” In *Building Successful Online Communities: Evidence-Based Social Design*, Robert E. Kraut and Paul Resnick, 125–178. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/8472.001.0001

Kumar, Sumit, and Raj Ratn Pranesh. 2021. “TweetBLM: A Hate Speech Dataset and Analysis of Black Lives Matter-related Microblogs on Twitter.” (https://arxiv.org/abs/2108.12521)

Livni, Ephrat. 2019. “Twitter, Facebook, and Insta Bans Send the Alt-Right to Gab and Telegram.” *qz*, May 12. Retrieved March 1, 2021 (https://qz.com/1617824/twitter-facebook-bans-send-alt-right-to-gab-and-telegram/).

Montanaro, Domenico. 2021. “Trump Teases Starting His Own Social Media Platform. Here's Why It'd Be Tough.” *NPR*, March 24. Retrieved June 1, 2021 (https://www.npr.org/2021/03/24/980436658/trump-teases-starting-his-own-social-media-platform-heres -why-itd-be-tough).

Munger, Kevin. 2017. “Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment.” *Political Behavior* 39(3): 629–49.

——. 2020. “Don’t @ Me: Experimentally Reducing Partisan Incivility on Twitter.” *Journal of Experimental Political Science* 8(2): 102–16. doi:10.1017/XPS.2020.14

Müller, Karsten, and Carlo Schwarz. 2018. “Fanning the Flames of Hate: Social Media and Hate Crime.” *Journal of the European Economic Association* 19(4): 2131–67.

——. 2020. “From Hashtag to Hate Crime: Twitter and Anti-Minority Sentiment.” *Available at SSRN 3149103* (https://ssrn.com/abstract=3149103).

Nagin, Daniel S. 1998. “Criminal Deterrence Research at the Outset of the Twenty-First Century.” *Crime and Justice* 23: 1–42.

Paluck, Elizabeth Levy, and Donald P. Green. 2009a. “Deference, Dissent, and Dispute Resolution: An Experimental Intervention Using Mass Media to Change Norms and Behavior in Rwanda.” *American Political Science Review* 103(4): 622–44.

——. 2009b. “Prejudice Reduction: What Works? A Review and Assessment of Research and Practice.” *Annual Review of Psychology* 60(1): 339–67.

Paternoster, Raymond. 1987. “The Deterrent Effect of the Perceived Certainty and Severity of Punishment: A Review of the Evidence and Issues.” *Justice Quarterly* 4(2): 173–217.

Pennycook, Gordon, Ziv Epstein, Mohsen Mosleh, Antonio A. Arechar, Dean Eckles, and David G. Rand. 2021. “Shifting Attention to Accuracy Can Reduce Misinformation Online.” *Nature* 592(7855): 590–95.

Peters, Jay. 2020. “Twitter Now Bans Dehumanizing Remarks Based on Age, Disability, and Disease.” *theverge*, March 5. Retrieved March 1, 2021 (https://www.theverge.com/2020/3/5/21166940/twitter-hate-speech-ban-age-disability-disease-dehumanize).

Pettigrew, Thomas F. 1998. “Intergroup Contact Ttheory.” *Annual Review of Psychology* 49(1): 65–85.

Rogers, Ronald W. 1975. “A Protection Motivation Theory of Fear Appeals and Attitude Change.” *Journal of Psychology* 91(1): 93–114.

Samii, Cyrus. 2013. “Perils or Promise of Ethnic Integration? Evidence from a Hard Case in Burundi.” *American Political Science Review* 107(3): 558–73.

Sherman, Lawrence W. 1993. “Defiance, Deterrence, and Irrelevance: A Theory of the Criminal Sanction.” *Journal of Research in Crime and Delinquency* 30(4): 445–73.

Siegel, Alexandra A., and Vivienne Badaan. 2020. “# No2Sectarianism: Experimental Approaches to Reducing Sectarian Hate Speech Online.” *American Political Science Review* 114(3): 837–55.

Silic, Mario, Dario Silic, and Goran Oblakovic. 2016. “Restrictive Deterrence: Impact of Warning Banner Messages on Repeated Low-Trust Software Use.” Presented at the 18th International Conference on Enterprise Information Systems (ICEIS 2016), April 25-28. http://doi.org/10.5220/0005831904350442

Simonovits, Gábor, Gabor Kezdi, and Peter Kardos. 2018. “Seeing the World through the Other's Eye: An Online Intervention Reducing Ethnic Prejudice.” *American Political Science Review* 112(1): 186–93.

https://doi.org/10.1017/S1537592721002589 Published online by Cambridge University Press

Spangler, Todd. 2020. "Reddit Finally Bans Hate Speech, Removes 2,000 Racist and Violent Forums Including The_Donald." *variety*, June 29. Retrieved March 1, 2021 (https://variety.com/2020/digital/news/reddit-bans-hate-speech-groups-removes-2000-subreddits-donald-trump-1234692898/).

Stafford, Mark C., and Mark Warr. 1993. "A Reconceptualization of General and Specific Deterrence." *Journal of Research in Crime and Delinquency* 30(2): 123–35.

Stockman, Mark, Robert Heile, and Anthony Rein. 2015. "An Open-Source Honeynet System to Study System Banner Message Effects on Hackers." In *Proceedings of the 4th annual ACM Conference on Research in Information Technology*, 19-2.

Takikawa, Hiroki, and Kikuko Nagayoshi. 2017. "Political Polarization in Social Media: Analysis of the "Twitter Political Field" in Japan." *2017 IEEE International Conference on Big Data (Big Data).* https://doi.org/10.1109/BigData41644.2017

Testa, Alexander, David Maimon, Bertrand Sobesto, and Michel Cukier. 2017. "Illegal Roaming and File Manipulation on Target Computers: Assessing the Effect of Sanction Threats on System Trespassers' Online Behaviors." *Criminology & Public Policy* 16(3): 689–726.

Wilson, Theodore, David Maimon, Bertrand Sobesto, and Michel Cukier. 2015. "The Effect of a Surveillance Banner in an Attacked Computer System: Additional Evidence for the Relevance of Restrictive Deterrence in Cyberspace." *Journal of Research in Crime and Delinquency* 52(6): 829–55.

Wogalter, Michael S. 2006. "Communication-Human Information Processing (C-HIP) Model." *Handbook of Warnings: Case Studies and* Analyses, 51–61. Boca Raton: CRC Press.

Ziems, Caleb, Bing He, Sandeep Soni, and Srijan Kumar. 2020. "Racism Is a Virus: Anti-Asian Hate and Counterhate in Social Media during the COVID-19 Crisis." arXiv preprint. (arXiv:2005.12423).

https://doi.org/10.1017/S1537592721002589 Published online by Cambridge University Press