European Sociological Review, 2018, Vol. 34, No. 3, 223–237
doi: 10.1093/esr/jcy005
Advance Access Publication Date: 19 March 2018
Original Article

Normative Change and Culture of Hate: An Experiment in Online Environments

Amalia Álvarez-Benjumea[1,2]* and Fabian Winter[1]

[1]Max Planck Research Group 'Mechanisms of Normative Change', Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, D-53113 Bonn, Germany and [2]University of Cologne, 50923 Köln, Germany

*Corresponding author. Email: alvarezbenjumea@coll.mpg.de

Submitted June 2017; revised February 2018; accepted February 2018

Abstract

We present an online experiment in which we investigate the impact of perceived social acceptability on online hate speech, and measure the causal effect of specific interventions. We compare two types of interventions: counter-speaking (informal verbal sanctions) and censoring (deleting hateful content). The interventions are based on the belief that individuals infer acceptability from the context, using previous actions as a source of normative information. The interventions are based on the two conceptualizations found in the literature: (i) what do others normally do, i.e. descriptive norms; and (ii) what happened to those who violated the norm, i.e. injunctive norms. Participants were significantly less likely to engage in hate speech when prior hate content had been moderately censored. Our results suggest that norm adherence in online conversations might, in fact, be motivated by descriptive norms rather than injunctive norms. With this work we present some of the first experimental evidence investigating the social determinants of hate speech in online communities. The results could advance the understanding of the micro-mechanisms that regulate hate speech. Also, such findings can guide future interventions in online communities that help prevent the spread of hate.

Introduction

The rise of online social interaction has opened the way for increased social participation. At the same time it has unlocked new gates to express hostility, making engagement harder for vulnerable groups, such as women, the LGBT community,[1] or other minority groups (Kennedy and Taylor, 2010; Mantilla, 2013). This behaviour is commonly referred to as hate speech. Hate speech is defined as speech intended to promote hatred on the basis of race, religion, ethnicity, or sexual orientation. It is closely related to other types of online antisocial behaviour, such as online harassment and trolling (Binns, 2012), since people who engage in these types of behaviour often make use of such methods. In this article, however, we will limit the analyses to hate speech, as we understand trolling as an umbrella term for different antisocial behaviours. We define hate speech as hostile behaviour and 'antagonism towards people' (Gagliardone *et al.*, 2015: p. 11) who are part of a stigmatized social group. The concept is, therefore, close to prejudice expression.[2]

Hate speech may cause fear (Hinduja and Patchin, 2007) and push people into withdrawing from the public debate, therefore harming free speech (Henson, Reyns, and Fisher, 2013) and contributing to a toxic online environment. Social platforms and organizations established to combat hate speech have recognized that online hateful content is increasingly common.[3] As a result, many governments and online media platforms have implemented diverse campaigns and interventions to tackle online hate speech.[4] Efforts against online hate often involve favouring counter-speaking (flagging, reporting, etc.) or censoring the hate content (Citron and Norton, 2011; Goodman and Cherubini, 2013). The theoretical and policymaking importance of these interventions has not yet been well understood.

We conducted a novel experiment to further our understanding of the underlying mechanism of hate speech. We tested whether decreasing social acceptability of hostile comments in a forum could prevent hate expression, and measured the causal effect of specific interventions. We used interventions designed to reduce hate speech in online environments: censoring hate content and counter-speech.[5] We designed an online forum and invited participants to join and engage in conversation on current social topics. We chose an online forum because online discussions are the basis of many social platforms on the Internet. Our experiment manipulates the comments participants could see before giving their own comments. The *censoring* treatment is a top-down approach that consists of censoring hate content and presenting an environment where prior hate comments are not observed. In the *counter-speaking* condition, the hate comments are presented with comments calling attention to the unacceptability of hate speech. The experiment was conducted with 180 subjects recruited from a crowdsourcing platform. We collected the comments from conversations in the forum and compared the level of hostility of the comments and instances of hate resulting across the conditions.

The interventions are based on the theoretical claim that social acceptability can be inferred from previous action. This claim is based on the observation that presenting a context where antisocial behaviour is common brings about more antisocial behaviour, such as littering, stealing, or jaywalking (Cialdini, Reno, and Kallgren, 1990; Keizer, Lindenberg, and Steg, 2008; Keuschnigg and Wolbring, 2015). A similar process has been found in online contexts, where prior troll comments affect the likelihood of subsequent trolling (Cheng *et al*., 2017). This cascading dynamic is linked to a process of spreading norm violations: people learn from each other which kind of behaviour is approved and which behaviour people are to expect in particular situations.

When people observe that others have violated a certain social norm, such as expressing hateful views, they are more likely to transgress it because they perceive the behaviour as socially accepted. The opposite should hold true: reducing the social acceptability of hateful behaviour online might reduce the willingness of individuals to engage in hate speech. This relationship between perceived social norms and prejudice expression in offline settings has been widely studied (Pettigrew, 1991; Paluck and Green, 2009). Highlighting a majority norm, or a perceived consensus against the expression of prejudice, reduces people's willingness to express prejudice (Stangor, Sechrist, and Jost, 2001; Crandall, Eshleman, and O'Brien, 2002; Crandall and Stangor, 2005; Shapiro and Neuberg, 2008). Parallel empirical research in online communities is still scarce, with some evidence of the effect of perceived norms on hate expression, such as the effect of promoting a norm through social sanction (Munger, 2016) or reminding participants of etiquette rules (Matias, 2016).

The experimental approach allows us to study the production of hate speech under very controlled conditions. Data from observational studies might present several sources of variation that make it difficult to disentangle different competing mechanisms. While users in those communities have already filtered content and self-selected themselves into contexts, we created and *randomly* assigned participants to those conditions to study their effects on the production of hate speech. Online experimentation introduces further advantages, such as increased anonymity—both among participants and towards the experimenter—and a recreation of the natural context where the behaviour of interest normally occurs, which increases ecological and external validity (Shadish, Cook, and Campbell, 2002; Rauhut and Winter, 2012).

Theory and Hypotheses

Social norms are shared rules that provide the standard of behaviour within a wide range of settings (Elster, 1989; Coleman, 1990a, 1990b; Hechter and Opp, 2001; Bicchieri, 2005), and the behaviour that people are to expect in particular situations.[6] Individuals are motivated to understand norms in their social context because they care about how they are perceived by others (DellaVigna *et al*., 2016), to avoid rejection (Cialdini and Goldstein, 2004), or to avoid further sanctioning (Hechter and Opp, 2001). Norms are usually not clearly determined and individuals rely on their subjective perceptions of norms (Tankard and Paluck, 2016); people use the behaviour of others as a source of information about social norms, and follow a norm conditional on their expectations about how others behave and how others believe one should behave in similar situations (Bicchieri, 2005).

The way we communicate is also regulated by social norms; in particular, individuals avoid publicly expressing views if they believe they are not popular in their social context (Cialdini and Trost, 1998; Cialdini and Goldstein, 2004; Bursztyn, Egorov and Fiorin, 2017). Likewise, prejudice is subject to similar normative influence.[7] There is evidence of a norm against public expression of hate in Europe (Ivarsaten, Blinder, and Ford, 2010; Blinder, Ford, and Ivarsaten, 2013), which makes an expression of prejudice more likely in a private than in a public context (Ford, 2008).[8] Because norms are interdependent, information about the behaviour of others is pivotal for normative change. For example, providing consensus information over negative stereotypes (Stangor *et al*., 2001) can reduce the adherence of people to prejudiced views. Events such as elections can have an effect, as they disclose information on the prevalence of certain opinions and induce changes in the perception of social acceptability. Bursztyn, Egorov and Fiorin (2017) argue that the 2016 election results in the United States causally increased individuals' perception of the social acceptability of anti-immigration views and 'their willingness to publicly express them'.

Observing the behaviour of others around us is a key source of information on established social norms (Bicchieri, 2005). Lab experiments show that prejudice expression can be reduced by manipulating 'normative acceptability of prejudice' (Blanchard *et al*., 1994, p. 362), showing consensus information over negative stereotypes (Stangor *et al*., 2001), or hearing others endorse an anti-prejudice norm (Zitek and Hebl, 2007). Further experiments found that individuals not only suppress prejudice expression but also are more likely to oppose discrimination immediately after hearing someone else do so (Cialdini and Trost, 1998).

Conveying Information about Appropriateness

We have argued that individuals infer acceptability from the context using the behaviour of other actors within it as a source of normative information. Previous literature has identified two sources of information: (i) what do others normally do, and (ii) what happened to those who violated the norm. This is the distinction between 'what normally happens', i.e. descriptive norms, and 'what others think one ought to do', i.e. injunctive norms (Cialdini *et al*., 1990; Cialdini and Goldstein, 2004; Bicchieri and Xiao, 2009). Descriptive norms act as coordination devices of 'normal behavior' (Bicchieri, 2005; Krupka and Weber, 2013), whereas injunctive norms act as an 'oughtness rule' (Hechter and Opp, 2001). Situational triggering cues can also increase the saliency of normative information and reduce ambiguity about the appropriateness of a certain type of behaviour. Actions that stand out, such as observing someone punishing, draw attention to the existing norm. The implications for online behaviour are straightforward. While observing prevalent online behaviour illustrates the descriptive norm, observing responses to those behaviours teaches the injunctive norms.

Building on the distinction between injunctive and descriptive norms, we operationalize the online setting in a way that allows us to study whether people learn about norms by observing them (descriptive norm mechanism) or by observing norm violations being sanctioned (injunctive norm mechanism). To do so, we adapt interventions designed to reduce hate speech in online environments: censoring hate content and letting peers verbally sanction it. Censoring hostile content biases the individual's perception of the prevalence of hate speech, i.e. the descriptive norm. If norms are followed because individuals perceive that a majority adheres to it, expectations are that people will not make use of hate speech. This mechanism predicts a direct relationship between the subject's action and what she observes others have done, thus:

Hypothesis 1a: Removing examples of hate speech in the online context, therefore decreasing its observed prevalence, will accentuate a descriptive norm and lead to less hostile content.

However, people making the choices are also heterogeneous, i.e. they require different amounts of social pressure to elicit a particular response. It is possible that merely deleting hate speech instances would not be a strong enough signal of an anti-hate norm for a majority of individuals. Thus, we have that:

Hypothesis 1b: Presenting only cases of friendly speech, therefore increasing its observing prevalence, will accentuate a descriptive norm and lead to less hostile content and fewer instances of hate.

Observing explicit counter-comments to hate content, i.e. verbal sanctions to a hateful comment, signals injunctive norms and clarifies the behaviour that is believed to be appropriate. We decided to use verbal sanctions because they fit naturally into an online conversation setting. Also, verbal sanctions, such those in online firestorms, are used as online normative enforcement (Rost, Stahel, and Frey, 2016), also against hate speech (Schieb and Preuss, 2016). If individuals need to see the consequences of behaviour to learn its appropriateness, then we have that:

Hypothesis 2: Observing verbal sanctions to previous examples of hate speech strongly signals the existence of the injunctive norm and will lead to less hostile content.

Experiment

Experimental Design

To test the hypotheses, we designed an online forum where participants could discuss current social topics. The online forum is designed to resemble an Internet forum.[9] Participants were invited to join the conversations and leave comments on topics portrayed in pictures. We collected their comments and later analysed them.

Pictures and topics were selected in a pre-experimental online survey ($N = 90$) from a list of 10 different social topics and 200 pictures.[10] We chose topics and pictures identified as controversial in the survey to ensure that all topics were, to some extent, subject to public debate. In total, nine pictures illustrating four topics were selected: three pictures on feminism, two pictures on LGBT rights, three pictures on refugees and multiculturalism, and one picture representing poverty. In a pre-experimental session, we made our forum available online and collected comments on the pictures. A team of three external raters classified a pool of 840 comments based on their level of hate speech into three categories: neutral, friendly, and hostile. The comments were later used to create the experimental conditions.

To test the effectiveness of censoring and counter-speaking, we modified the comments thread in the discussion forum among treatments, while maintaining the order in which the pictures were presented. All participants were presented with the same sequence of nine threads discussing each picture. Because the sequence of the pictures is identical between participants, we do not consider the pictures part of the experimental condition.

All comments used to create the experiment came from the pre-experimental session, including the comments used as peer-sanctions in the counter-speaking treatment. The complete experiment timeline is described in Figure A7 in Appendix A.

Experimental Treatments

We implemented four different treatments: baseline, censored, extremely censored, and counter-speaking. The treatments vary in the comments composing the discussion threads in the forum. In the baseline condition, participants could see a balanced mix of two friendly, two neutral, and two hostile comments.

Censored conditions

We implemented two versions of the censored conditions to test *Hypotheses 1a* and *1b*, respectively: censored and extremely censored. Both conditions were designed to highlight a descriptive norm against hate expression. In the censored condition, we deleted prior hate content and presented participants only with friendly and neutral comments. In the extremely censored condition, we presented only friendly comments. Information on whether comments had been deleted was not displayed.

Counter-speaking condition

In the counter-speaking condition, the hostile comments were presented with replies highlighting the unacceptability of hostile opinions (e.g. '@[user] this is a prejudiced judgment'). The replies are verbal sanctions that make an injunctive norm salient. The replies were collected from participants in a pre-experimental session. A total of 117 verbal sanctions to hostile comments were collected.

The exact comments to be shown in the discussion were automatically selected from a database for each participant in the experiment.[11] Table 1 summarizes the number of comments in the different experimental conditions.

Data Collection

The experiment was conducted entirely online and participants were recruited from a crowdsourcing Internet marketplace.[12] The experiment was conducted entirely in German, and the sample was restricted to German residents. Although we did not directly ask participants for their demographic characteristics, the subjects were selected from a population with the characteristics

Table 1. Summary of the content of the forum in the different treatments

Treatment	Summary of forum content
Baseline	Six comments: two friendly, two neutral, and two hostile
Counter-speaking	Six comments: one friendly, one neutral, and two hostile plus two sanctions
Censored	Four comments: two friendly and two neutral
Extremely censored	Three comments: all friendly

depicted in Table 2. The sample is obviously more diverse than the traditional convenience sample of students.

Participants were compensated with a fixed amount of 3 euros. To avoid demand effects, the participants were told that they were taking part in an experimental study, but not told the purpose of the experiment. Links were posted in the recruiting platform, and upon acceptance participants were redirected to our own online forum. Participants were randomly allocated between the conditions and asked to join the discussion forums. Each participant was then showed the introduction page explaining the nature of the task. Participants could abandon the experiment at every stage of the experiment just by closing the browser. At the beginning of the experiment, they were given a randomly generated neutral user name and avatar. Every participant was consecutively presented with the nine discussions and asked to leave a comment at the bottom of each thread. Giving a comment was mandatory in each of the nine discussions. Navigation throughout the online forum was always forward. It was not possible to go back to previous discussions once a comment had been sent. When the experiment was completed the participants were given a code to claim the payment for participating in the experiment.

A total of 180 participants were recruited to take part in the forum. Participants spent an average of 10 min in the forum (Table 3). We collected a total of 1,585 comments, of which 116 were invalid.[13] The comments were evenly distributed among the pictures, with a maximum of 180 and a minimum of 174 per picture. Participants could not see what other participants immediately before them had commented, but only the comments we had previously selected to create the different conditions. This ensured that individual observations were independent.

Table 2. Sociodemographic characteristics of the population from where subjects were recruited

Gender	
Women	55%
Age (years)	
18–24	28%
25–34	42%
35–44	17%
>45	13%
Employment status	
Student	29%
Employee	26
Self-employed	15
Other	20
N.S	10

Measurement of Variables and Operationalization

We evaluated the comments in two ways: we assigned them a score, using a nine-point scale measuring hostility; and we identified those that were clear violations of an anti-hate speech norm. The score tries to encompass a broad definition of hate speech in terms of 'tolerance, civility, and respect to others' (Gagliardone *et al.*, 2015: p. 15). This measure is related to the notion of hate speech based on social norms of polite expression. The second measure refers more to a notion of hate speech similar to those found in legislations and international agreements, such as the policy recommendations of the European Commission against Racism and Intolerance (ECRI, 2015), which name the most egregious forms of hate speech.

Hate speech score

The first outcome of interest is the change in the level of hostility displayed by the subjects in the different treatments compared to the baseline group. Thus, the collected comments were classified following a hate speech score by three external raters blind to the experimental conditions. Raters were provided with each comment and the question: 'Is the comment friendly or hostile towards 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 a score of 1 are very friendly in language and express a positive opinion (e.g. Comment 110: 'Very brave, I find it great and refreshing. I find despising homosexuals generally bad', User 84, *Schwarzbeere*. Retrieved from a thread on LGBT rights), whereas a score of 9 normally implies aggression (e.g. Comment 1,029: 'Gay guys are the last thing I would tolerate, especially in public', User 112. Retrieved from a thread in LGBT rights).

We opted for a continuous measure instead of a binary classification because binary classifications of hate

Table 3. Number of participants and valid comments per treatment

Treatment	Subjects	Comments
Baseline	47	375
Counter-speaking	45	373
Censored	46	377
Extremely censored	42	344
Total	**180**	**1,469**

speech have been found not to be very reliable (Ross *et al.*, 2016). Inter-rater reliability of our scale is relatively high (Krippendorf's $\alpha = 0.69$).[14] Thus, we averaged the three scores to construct a hate speech score. The continuous score allows us to study subtle variations and serves as the main variable of interest in the study.

Hate speech indicators

We identified various items in the literature that are consistently considered instances of hate speech: (i) contains negative stereotypes, (ii) uses racial slurs, (3) contains words that are insulting, belittling, or diminishing, (iv) calls for violence, threat, or discrimination, (v) uses sexual slurs, and (vi) sexual orientation/gender used to ridicule or stigmatize.

These items are based on guidelines on how to detect online hate speech, published by UNESCO (Gagliardone *et al.*, 2015), as well as the ECRI general policy recommendation on combating hate speech (2015).[15] Comments were labelled as violations of the anti-hate-speech norm if they contain one of the listed indicators. This measure was created with the intention of having a more systematic classification of the norm violations, which could be used for robustness checks.

The two variables measure different things, which can lead to mismatches. Nevertheless, they are closely related and, as the value of the score increases, the probability of a comment being labelled as hateful also increases.[16] The following comment is a typical example of hate content in the forum with a score of 8.66. Comment 159: 'Refugee crisis. They can continue walking away from Europe. They are not just war refugees, 90 per cent are nothing but social parasites who can do whatever they want here'. (User 171, *Springfrosch*. Retrieved from a thread on refugees/multiculturalism. The original comment is in German.). The comment was also marked as containing Items 1 and 3 by the three raters and therefore classified as a norm violation. More examples of comments can be found in Appendix A6.

Data and Results

Average Levels of Hate Speech

We begin this section by analysing the hate speech score. The mean differences in hate speech score by treatment across topics are displayed in Figure 1. The mean hate speech score is reduced in all treatments compared to the baseline treatment for all topics except poverty.

We estimated a random intercept regression model (Snijders, 2011; Judd, Westfall, and Kenny, 2017) with two random factors, subjects and pictures, and hate speech score as the dependent variable to assess the ability of the treatments to reduce average levels of hostility (see Table 4).[17] The main explanatory variables are the treatments (Model 1), but we also included terms for the different topics (Model 2), and a term for each combination of treatment and topic (Model 3). Although the

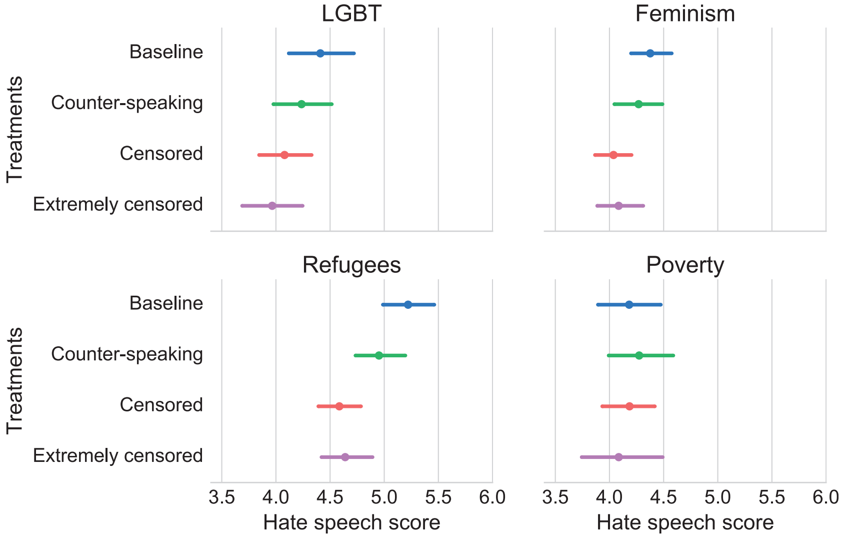

Figure 1. Treatment differences in mean hate speech score across topics (observations=1,469). Error bars at the 95% confidence interval.

Table 4. Results from multilevel random models of hate speech score

	Model (1)	Model (2)	Model (3)
Main effects			
Intercept	4.63 (0.17)**	4.41 (0.35)**	4.20 (0.37)**
Counter-speaking	−0.14 (0.16)	−0.14 (0.16)	
Censored	−0.39 (0.15)*	−0.39 (0.15)*	
Extremely censored	−0.40 (0.16)*	−0.40 (0.16)*	
LGBT		−0.00 (0.41)	0.22 (0.44)
Refugees/multiculturalism		0.61 (0.39)	0.97 (0.41)*
Feminism		0.03 (0.39)	0.19 (0.41)
Interaction effects			
Poverty*counter-speaking			0.07 (0.24)
LGBT*counter-speaking			−0.16 (0.20)
Refugees*counter-speaking			−0.26 (0.19)
Feminism*counter-speaking			−0.10 (0.18)
Poverty*censored			−0.02 (0.24)
LGBT*censored			−0.33 (0.20)†
Refugees*censored			−0.64 (0.19)**
Feminism*censored			−0.33 (0.18)†
Poverty*extremely			−0.12 (0.25)
LGBT*extremely			−0.45 (0.21)*
Refugees*extremely			−0.60 (0.19)**
Feminism*extremely			−0.28 (0.19)
Random parts			
Residual variance	0.90	0.90	0.90
Groups: Subjects	180	180	180
Var: Subjects	0.44	0.44	0.44
Groups: Pictures	9	9	9
Var: Pictures	0.15	0.11	0.11
ICC: Subjects	0.30	0.30	0.30
ICC: Pictures	0.10	0.07	0.07
AIC	4,345.59	4,347.50	4,371.48
BIC	4,382.64	4,400.42	4,472.04
Observations	1,469	1,469	1,469

Notes: Linear mixed model fit by REML. Fixed-effects estimates (Top) and variance–covariance estimates (Bottom) for models of hate speech score. Model 1 shows main effects of treatments, Model 2 shows main effects of topics, and Model 3 shows the interaction between treatments and topics after controlling for topic main effects. The table lists mean regression coefficient estimates with standard errors in parentheses and *P*-values calculated based on Satterthwaite's approximations. AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; REML = Residual Maximum Likelihood; var = variance.
Significance levels: ***$P < 0.000$, **$P < 0.01$, *$P < 0.05$, †$P < 0.1$, for two-sided tests.

effects of the treatments by topics were not part of our original research question, including them allows us to ensure that the effect is not driven by just a single topic.

We computed the following models:

$$Y_{ijk} = \beta_0 + \beta_1 Treatment_{ij} + u_i + v_j + \epsilon_{ijk} \quad (1)$$

$$Y_{ijk} = \beta_0 + \beta_1 Treatment_{ij} + \beta_2 Topic_{ijk} + u_i + v_j + \epsilon_{ijk} \quad (2)$$

$$Y_{ijk} = \beta_0 + \beta_1 Topic_{ijk} + \beta_2 (Treatment\ _{ij}x\ \ Topic_{ijk}) + u_i + v_j + \varepsilon_{ijk} \quad (3)$$

where $u_i{}^{\sim}N(0, \sigma_u)$ and $v_j{}^{\sim}N(0, \sigma_j)$

The first model in Table 4 shows all treatments compared to the baseline condition. The censored and extremely censored conditions significantly reduced hostility by 0.39 and 0.40 scale points, respectively. These results show support for the descriptive norm mechanism suggested in *Hypotheses 1a* and *1b*, although extremely censoring does not have an additional effect on the mean score, and the magnitude of the reduction is the same for both treatments. In the counter-speaking condition, the score is reduced by 0.14 points compared to the baseline treatment, but this reduction is not significant. These results show no support for *Hypothesis 2*. The second model in Table 4 adds topics as predictors, using poverty as the reference category. After controlling for the topics, the effect of the experimental predictors persists. Comments on refugees/

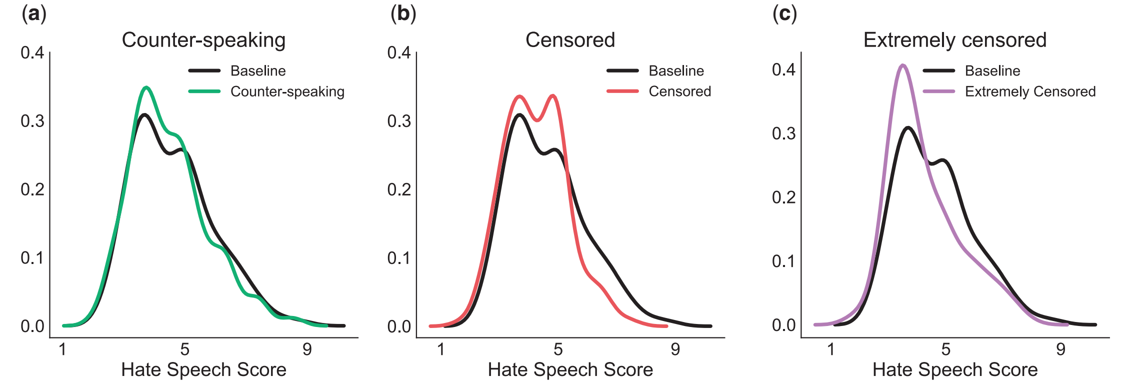

Figure 2. Density estimates of average hate speech score in the counter-speaking treatment, the censored condition, and the extremely censored condition, compared to baseline treatment. Size of bins selected using Freedman–Diaconis rule. All treatments are compared to the baseline group ($N = 1,469$). The graph depicts the 1–9 score scale: scores on the left correspond to friendly speech, while scores on the right correspond to more hostile language.

multiculturalism threads are more hostile on average (0.61 scale points), while the rest of topics obtained similar levels of hostility to poverty.

The following comments, retrieved from a thread on transgender issues, illustrate what a change from 5 to 6 in the score looks like. Comment 268: ‘Much more important than the question of man, woman or transgender seems to me the question of why anyone goes to the military’. (User 81, *Halbmond*. Score of 5). Comment 209: ‘I am confused’. (User 180, *Kekskuchen*. Score of 5, 33). Comment 317: ‘I would claim to have chosen the wrong clothes in the wardrobe in the morning’. (User 25, *Wintergrün*. Score of 6).

Model 3 shows the effect of the treatments for each topic compared to their respective baseline levels. The treatments’ main effects are deliberately not included in the model. This way the effect of the treatments is shown for each topic specifically. Because poverty is used as the reference category, the intercept represents the estimate for poverty in the baseline treatment. Although the magnitude and significance of the effect differ between topics, the treatments consistently reduce the score as shown by the negative coefficients. In the case of the counter-speaking treatment, and in line with Model 1, this reduction is not significant for any topic. Similarly, none of the terms for the interaction of the treatments with poverty is significant. The treatment effect is larger in threads discussing pictures portraying refugees/multiculturalism, and both censoring and extremely censoring reduce the score in more than half point in this topic. These results should be interpreted with caution, since the effects of the topics are not part of the original research questions, and we do not have enough statistical power to test the assumption of a larger effect in threads on refugees/multiculturalism.

Distribution of the Hate Speech Score

Figure 2 displays the distribution of the hate speech score of each treatment compared to the baseline. Extreme comments, both extremely hateful and extremely friendly, are rare, with a majority of comments classified as neutral. The distribution of the hate speech score in the baseline and the counter-speaking conditions are similar. In contrast, in both censored treatments the distributions are skewed to the left, which means that comments were less hostile on average (for both treatments compared to the baseline a Kolmogorov–Smirnov test of equality of the distributions yields $P < 0.001$).

Next, we analyse displays of hostile comments. We define hostile comments as those with a 7 or more in the score. Hostile comments are relatively uncommon ($N = 61$). Of the comments in baseline treatment, 24 were hostile (5.88 per cent) compared to 4 comments in the censored condition (0.99 per cent). The reduction is significant [χ^2 (1,814) = 14.94, $P < 0.001$].[18] There are 18 hostile comments in the counter-speaking treatment (4.44 per cent) and 15 in the extremely censored treatment (4.10 per cent). None of them significantly reduces extremely hostile comments. Similar results are obtained using a quantile regression. We computed the treatment effect in the 0.75, 0.90, 0.95, and 0.99 quantiles of the hate score distribution. The censored condition significantly reduces hate in the 0.75, 0.90, 0.95, and 0.99 quantiles, whereas a significant effect of the extremely censored condition is found only at the 0.75 and 0.99 quantiles. The treatment effect of the censored condition is also larger in the higher quantiles than the mean effect, e.g. a reduction of 1.33

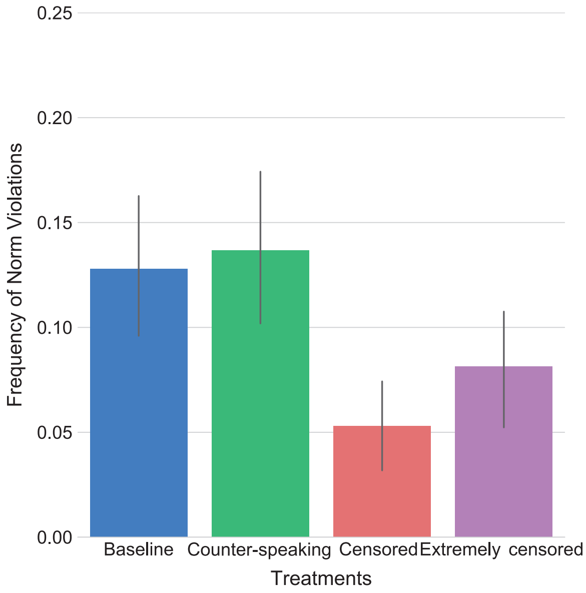

Figure 3. Proportion of comments that were labeled as hate speech across treatments (N = 1,469). Error bars at 95% confidence interval.

Table 5. Total number of hostile comments per participant that made at least one hostile comment

	Total number of hostile comments per participant						
Treatments	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$	$n = 7$
Baseline	6	0	4	0	0	1	0
Counter-speaking	7	3	0	0	1	0	0
Censored	2	1	0	0	0	0	0
Extremely censored	6	1	0	0	0	0	1
Total	**21**	**5**	**4**	**0**	**1**	**1**	**1**

scores points in the 0.99 quantile compared to the baseline (see the Appendix B). Of the total number of participants, 33 left a comment that was classified as hostile. Most participants that left a hostile comment left one or two in total. The maximum number of hostile comments left by one unique participant is seven. Table 5 shows the distribution of the total number of hostile comments made by participants that made at least one hostile comment.

In the extremely censored condition, there are more comments with hostile scores than in the censored condition [χ^2 (1,721) = 7.63, P <0.01]; even though participants shift their tone (from neutral to slightly friendly), there is an upturn of hostile language compared to the censored condition. This upturn effect is not robust if we take into account the nested structure of the comments, that is, it disappears when we analyse the distribution of hostile comments from the participants' perspective (see Appendix B).

Analysis of the Norm Violations

In addition to the hate score, we analysed comments that were classified as a norm violation according to our hate speech indicators, that is, comments that are regarded as uncivil.[19] In the analysis, only comments that were classified as a norm violation by two or three of the raters (majority rule) were used (N = 147).

We tested for differences in the frequency of uncivil comments among the different treatments, using a multilevel logistic model with a random effect for participants (Table 6).[20]

The predicted probability of observing a norm violation in the baseline treatment is 0.16. This probability is reduced to 0.06 in the censored condition and 0.10 in the extremely censored condition. Nevertheless, only the censored condition presents a significant reduction compared to the baseline condition. Again, we find support for the

Table 6. Results from a multilevel logistic model of the probability of a norm violation

	Model (1)
Fixed parts	
Intercept	−2.53 (0.30)***
Counter-speaking	0.19 (0.38)
Censored	−1.00 (0.42)*
Extremely censored	−0.50 (0.41)
Random parts	
Groups: Subjects	180
Var: Subjects	1.57
Groups: Pictures	9
Var: Pictures	0.05
AIC	892.46
BIC	924.22
Observations	1,469

Notes: Generalized linear mixed model fit by maximum likelihood (Laplace approximation). Fixed-effects estimates (Top) and variance–covariance estimates (Bottom) for models of norm violations. The table lists logistic regression coefficient estimates with standard errors in parentheses and *P*-values calculated based on Satterthwaite's approximations. AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; var = variance.

Significance levels: ***$P < 0.000$, **$P < 0.01$, *$P < 0.05$, $^{\dagger}P < 0.1$, for two-sided tests.

descriptive mechanism suggested in *Hypothesis 1a*. We found no significant differences in the number of norm violations by topic. Table 7 shows the distribution of uncivil comments made by participants that made at least one.

Limitations of the Study

There are three potential limitations concerning the external validity, the generalizability of our results, and the statistical inference. First, the static nature of the forum prevents people from engaging in repeated interaction, which departs from normal dynamics in online forums. The lack of interaction can be important to explain the failure of the injunctive mechanism to reduce hostility in the forum. If, for instance, the commenter expects their comments to be counter-commented, they might be more hesitant to post hateful content.

Secondly, our sample of participants is limited to online workers, whose characteristics may vary from the general population, thus limiting generalizability of the results. Online workers might differ from the average user of the Internet in their inclination to post hateful comments. This point is not limited to online labour markets but also applies, for example, to left-leaning or right-leaning websites. Because participation in the experiment was anonymous and voluntary, we believe that their motivation to express hostility should not differ from motivations of the general population of online forum commenters. Although we have no reason to assume that the particular treatment effects are qualitatively changed by our sampling strategy, the results in this article should not be interpreted as prevalence estimates of hate speech. We acknowledge that our treatments might have different effects for different people, i.e. no effect for those with a strong ideological leaning. Our data do not allow us to test this hypothesis. From a practical point of view, this would mean that providers of online platforms would have to apply very different policies for different people, which is normally not the case.

Finally, the limited size of the sample poses some problems to the statistical inference, especially when analyzing rare events as hostile comments. A bigger sample would be helpful, and this could be collected, for instance, from existing websites and social platforms. These data would be observational, with the endogeneity problem that goes along with it. In contrast, our data are collected in a controlled experimental environment, and therefore allow for a proper identification of the treatment effect.

Conclusion

The widespread use of social media has become a reality in an ever more densely connected world. One of the biggest social challenges regarding social media is to tackle the hateful speech present in online discussions because it can prevent minority groups from joining conversations and expressing their opinions. We introduced

Table 7. Total number of uncivil comments per participant that made at least one

	Total number of uncivil comments per participant						
Treatments	*n* = 1	*n* = 2	*n* = 3	*n* = 4	*n* = 5	*n* = 6	*n* = 7
Baseline	10	3	3	3	1	1	0
Counter-speaking	10	9	4	1	0	0	1
Censored	7	5	1	0	0	0	0
Extremely censored	16	0	1	1	1	0	0
Total	**43**	**17**	**9**	**5**	**2**	**1**	**1**

an original randomized experiment to test whether reducing the perceived acceptability of hostility decreases the prevalence of online hate speech. We used specific interventions that used previous actions of others as a source of norm-relevant information.

In line with our first hypothesis (*H1a*), we find that moderately censoring hate content reduces the occurrence of further hate comments. Participants were less likely to make use of hostile speech when they were presented with an environment in which previous extreme hate content had been censored. Our results suggest that people respond to cues, in the online context, which signal social acceptability. They do so even when others are unknown and participants remain anonymous, and in the absence of direct punishment. The empirical results do not fully support our second hypothesis (*H1b*). The general tone of the comments became friendlier by applying an extreme version of censoring, but the effect is similar to moderately censoring. Moreover, there are significantly more extreme hostile comments in the extreme censoring conditions and then in the censored condition, which hints at a polarization of opinions. Our intuition is that this result might indicate either a reactance effect.[21]

Both experimental manipulations were more effective for the threads on refugees or multiculturalism. This differential of effects is related to the level of public debate on the topic: a lot of public debate of a topic may increase the salience of the norm (e.g. people might have previously observed that extreme opinions have been sanctioned). However, the censored conditions also allow for potential competing mechanisms such as mimicking previous comments. Nevertheless, the high prevalence of hate comments in the extremely censored condition indicates that people do not merely imitate observed behaviour, but they interpret the actions of others as contextual cues.

The counter-speaking condition meant to test *Hypothesis 2* had no significant effect. A potential explanation is that the verbal sanctions, i.e. the counter-comments, might add ambiguity about the norm (the verbal sanctions are, essentially, hostile comments themselves) by putting descriptive and injunctive information in conflict. In ambivalent situations like this, in which more than one norm may apply, individuals may interpret the situation in a way that favours them (Bicchieri and Xiao, 2009; Bicchieri and Chavez, 2013, Winter, Rauhut and Miller, 2017).

Our findings contribute to the sociological literature in social norms by raising the question of whether descriptive norms might, in some settings, be more effective than sanctions at preventing antisocial behaviour. Our results suggest that normative behaviour in online conversations might, in fact, be motivated by descriptive norms rather than injunctive norms. This is a surprising effect, given the results from previous research on social norms that pointed to a large effect of sanctions on normative behaviour (Heckathorn, 1988; Coleman, 1990a; Voss, 2001). Lab experiments on social norms show similar findings. For example, Fehr and Gächter (2000) conclude that punishment is far more effective than mere suggestions of maintaining a cooperation norm in a lab experiment. Furthermore, when the effects of injunctive and descriptive norms have been tested together, they do not significantly differ from each other (Krupka and Weber, 2009). Nevertheless, this result should be taken with precaution, since of injunctive information might be weakened by the lack of interaction.

The experimental nature of the study allows us to exclude potential confounding factors that can substantially bias the analysis of observational data. The randomization of subjects between conditions eliminates selection effects, e.g. hateful commenters joining only hateful discussions, and the anonymity in the forum prevents the occurrence of group identification processes. This study overcomes the identification problems that often arise from estimating the effect of normative influence.

This project is a step forward in the empirical research on online hate speech. First, we show that the observed pattern of behaviour acts as a situational normative cue in online environments. Secondly, our findings point to a larger effect of descriptive norms—defined as frequent behaviour—on reducing hate speech. Finally, we provide a reliable empirical test of censoring and counter-speaking as interventions, and show that moderately censoring hate content is sufficient to reduce uncivil comments. We believe the results in this study can support the design of online platforms that help reduce the incidence of hate speech in cases where it is undesirable and maintain an open online environment.

Nevertheless, we do not have data on the obvious trade-off between censoring and free speech; hence, our article does not represent a position on whether censoring hate content is necessarily socially beneficial. We consider that a social norm intervention (Tankard and Paluck, 2016) is a good approach to address online hate speech, whose presence is not necessarily considered unlawful, but often regarded as undesired.

Notes

[1] Lesbian, gay, bisexual, and transgender.

[2] We will use the terms hate, hostility, and prejudice interchangeably in the text.

3 UN Human Rights Council Special Rapporteur on Minority Issues (HRC, 2015) or Council of Europe, Mapping study on projects against hate speech online (15 April 2012). For some statistics, see Hate Base (http://hatebase.org).

4 Concerns about hate speech and violence can be linked to responses at various levels. Digital platforms, for instance, allow for different responses. In many cases platforms present some type of moderation process (Goodman and Cherubini, 2013). Community guidelines, such as in Facebook (https://www.facebook.com/communitystandards) and Youtube (https://www.youtube.com/yt/policyandsafety/communityguidelines.html), are also common. International initiatives to keep track of hate speech across networks have also emerged, such as HateBase and Fight against Hate (Gagliardone *et al.*, 2015). At the national level, countries like Germany have made huge efforts to combat online hate speech. On June 2017, Germany approved a law, the Netzwerkdurchsetzungsgesetz, which requires social media sites to remove all hate and extremist content (Bundesgesetzblatt 2017 Teil I Nr. 61, 07.09.2017, 3352-3355)

5 Different interventions have been discussed in the literature. Goodman and Cherubini (2013) refer to pre- and post-moderation of content as a strategy for creating better conversations online. Kraut *et al.*, (2012) discuss evidence-based recommendations to design better online platforms. Among the strategies presented, using descriptive norms to tackle online hate speech is discussed (Kraut *et al.*, 2012: p.13). Furthermore, Schieb and Preuss (2016) present empirical evidence on counter-speech as a measure for tackling hate speech on online platforms such as Facebook. We use this discussion as motivation for our treatments and construct them as ideal types of existing interventions, which help us to identify clean treatment effects. For evidence-based general recommendations on how to design online communities, please see Kraut *et al.* (2012) and Goodman and Cherubini (2013).

6 Social norms might take the form of quick quasi-automatic answers or completely developed actions, since they are often grounded in “scripted sequences of behavior” (Bicchieri and McNally, 2016: 2).

7 Allport (1979) and Sherif and Sherif (1953) wrote seminal texts arguing that the majority of prejudiced attitudes arose from conformity to social normative expectations. Sherif and Sherif (1953) describe the development of prejudice-expression norms as the result of the pressure that the group places on individuals to conform to the group norms.

8 Rejection of public expression of prejudice has been generally increasing in the past decades in many western societies (Pettigrew, 1958; Duckitt, 1992), although differences between countries are broad. For example, European countries tend to have a stricter view of what can be considered hate speech, whereas in North America more weight is given to free speech (Pettigrew, 1958; Dovidio and Gaertner, 1986; Duckitt, 1992).

9 The forum was created using Otree (Chen, Schonger, and Wickens, 2016), a software platform for economics experiments.

10 Pictures were previously collected from online media. We used Twitter and Google images to collect the pictures using a set of keywords (for the list of keywords, see Appendix A).

11 As shown in Table 1, the number of comments differs by treatment. One might argue that the different number of comments could have an impact on the level of hostility and therefore be a confounder of the treatment effect. For example, fewer comments might discourage participants to comment. The decision to vary the number of comments was a design choice to keep the amount of friendly content more or less equal between treatments, and to avoid suspicious designs. Nevertheless, no traces of discouragement effect due to a low number of comments were found, since the number of invalid comments was evenly distributed among treatments. An information reduction effect does not seem probable, since the condition with the less displayed information is not the one with less participant-generated hostile content.

12 We recruited participants from Clickworker (www.clickworker.com). The advantage of using this platform was that we could prevent subjects from participating more than once in our experiment, a widespread problem of online experiments.

13 The number of invalid comments is evenly distributed among the treatments: 33 in the baseline, 29 in the censored, 22 in the extremely censored, and 32 in the counter-speaking treatment. A comment is considered invalid when it is unintelligible. We ran an analysis excluding the comments of those who failed to leave nine comments ($N = 6$), and the results did not change.

14 We computed different measures of inter-rater agreement and reliability such as intra-class

correlation (ICC = 0.704). We chose Krippendorf's α to assess inter-rater reliability. This measurement is commonly used by researchers in content analysis (Krippendorf, 2004), and it is well suited to handling missing data, as well as specially recommended for cases with more than two raters. The level of agreement differs between the topics. The maximum level of agreement is found in refugees/multiculturalism (α = 0.71) and the lowest in LGBT (α = 0.58).

15 'Considering that hate speech is to be understood for the purpose of the present General Policy Recommendation as the advocacy, promotion or incitement, in any form, of the denigration, hatred or vilification of a person or group of persons, as well as any harassment, insult, negative stereotyping, stigmatization or threat in respect of such a person or group of persons and the justification of all the preceding types of expression, on the ground of "race", colour, descent, national or ethnic origin, age, disability, language, religion or belief, sex, gender, gender identity, sexual orientation and other personal characteristics or status'. (ECRI, 2016: p. 3)

16 The predicted probabilities of a comment being classified as hateful by the three raters increase from 0.054 at a score of 6 to 0.98 at a score of 9. Only the 5.5 per cent of comments containing an item from the list has an average score of 5 or lower.

17 Subjects were asked to leave nine comments in nine different pictures; hence the comments are clustered both within subjects and pictures. The models with one random level and two random levels were tested using analysis of variance. Both random levels are significant. The magnitude of the ICC estimate, i.e., variance accounted for by between-subject differences, suggests that variability between subjects is very high and should be taken into account in all analysis.

18 Our findings are robust to the different inference methods as displayed in Table A2 in Appendix B. The effect of the censored treatment is robust to the removal of influential individuals. These analyses are available upon request.

19 The analysis made on the frequency of hate speech comments across treatments should be made very carefully because the rate of agreement between the raters is low (Krippendorf's α = 0.40). Ross et al. (2016) found that inter-rater agreement for binary classification of tweets as hateful or not hateful is very low. Our rate of agreement is, in any case, higher than theirs.

20 Our results are also robust to different testing methods (see Table A3 in Appendix B), and to the removal of the most influential individuals (analyses available upon request).

21 Reactance appears when an individual facing a persuasive message reacts by engaging in the proscribed behavior (Burgoon *et al.*, 2002).

Supplementary Data

Supplementary data are available at *ESR* online.

Funding

Funding by the Max Planck Society is gratefully acknowledged.

References

Allport, G. W. (1979). *The Nature of Prejudice*. New York: Basic Books.

Bicchieri, C. (2005). *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge: Cambridge University Press.

Bicchieri, C. and Chavez, A. K. (2013). Norm manipulation, Norm evasion: experimental evidence. *Economics and Philosophy*, **29**, 175–198.

Bicchieri, C. and McNally, P. (2016). *Shrieking Sirens. Schemata, Scripts, and Social Norms: How Change Occurs (No. 0005)*. Wprking paper from Philosophy, Politics and Economics. Pennsylvania, PA: University of Pennsylvania.

Bicchieri, C. and Xiao, E. (2009). Do the right thing: but only if others do so. *Journal of Behavioral Decision Making*, **22**, 191–208.

Binns, A. (2012). DON'T FEED THE TROLLS! Managing troublemakers in magazines' online communities. *Journalism Practice*, **6**, 547–562.

Blanchard, F. A. *et al.* (1994). Condemning and condoning racism: a social context approach to interracial settings. *Journal of Applied Psychology*, **79**, 993.

Blinder, S., Ford, R. and Ivarsaten, E. (2013). The better angels of our nature: how the antiprejudice norm affects policy and party preferences in Great Britain and Germany. *American Journal of Political Science*, **57**, 841–857.

Burgoon, M. *et al.* (2002). Revisiting the Theory of Psychological Reactance. *The Persuasion Handbook*. Thousand Oaks, CA: Sage, pp. 213–232.

Bursztyn, L., Egorov, G. and Fiorin, S. (2017). *From Extreme to Mainstream: How Social Norms Unravel*. Working Paper 23415. National Bureau of Economic Research, available from: http://www.nber.org/papers/w23415.

Cheng, J. *et al.* (2017). Anyone can become a troll: Causes of trolling behavior in online discussions. arXiv preprint arXiv: 1702.01119.

Chen, D. L., Schonger, M. and Wickens, C. (2016). oTree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, **9**, 88–97.

Cialdini, R. B. and Goldstein, N. J. (2004). Social influence: compliance and conformity. *Annual Review of Psychology*, **55**, 591–621.

Cialdini, R. B., Reno, R. R. and Kallgren, C. A. (1990). A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, **58**, 1015.

Cialdini, R. B. and Trost, M. R. (1998). Social influence: social norms, conformity, and compliance. In Gilbert, D. T., Fiske, S. T. and Lindzey, G. (Eds.), *The Handbook of Social Psychology*. Vol. **2**, 4th edn. New York: McGraw-Hill, pp. 151–192.

Citron, D. K. and Norton, H. (2011). Intermediaries and hate speech: fostering digital citizenship for our information age. *Boston University Law Rev*, **91**, 1435.

Coleman, J. S. (1990a). The emergence of norms. In Hechter, M., Opp, K. -D., and Wippler, R. (Eds.), *Social Institutions: Their Emergence, Maintenance, and Effects*. New York: Aldine de Gruyter, pp. 35–39.

Coleman, J. S. (1990b). *Foundations of Social Theory*. Cambridge, MA: The Belknap Press of Harvard University.

Crandall, C. S., Eshleman, A. and O'Brien, L. (2002). Social norms and the expression and suppression of prejudice: the struggle for internalization. *Journal of Personality and Social Psychology*, **82**, 359.

Crandall, C. and Stangor, C. (2005). Conformity and prejudice. In Dovidio, J. F., Glic, P. and Rudman, L. A. (Eds.), *On the Nature of Prejudice: Fifty Years after Allport*. Malden, MA: Blackwell, pp. 295–309.

DellaVigna, S. *et al.* (2016). Voting to tell others. *The Review of Economic Studies*, **84**, 143–181.

Dovidio, J. F. and Gaertner, S. L. (1986). *Prejudice, Discrimination, and Racism: Historical Trends and Contemporary Approaches*. San Diego, CA: Academic Press.

Duckitt, J. H. (1992). Psychology and prejudice: a historical analysis and integrative framework. *American Psychologist*, **47**, 1182.

Elster, J. (1989). Social norms and economic theory. *The Journal of Economic Perspectives*, **3**, 99–117.

European Commission against Racism and Intolerance. (2016). *Recommendation No. 15 on Combating Hate Speech, Adopted on December 2015 (General Policy Recommendation)*. Strasbourg: Council of Europe.

Fehr, E. and Gächter, S. (2000). Cooperation and punishment in public goods experiments. *The American Economic Review*, **90**, 980–994.

Ford, R. (2008). Is racial prejudice declining in Britain?. *The British Journal of Sociology*, **59**, 609–636.

Gagliardone, I. *et al.* (2015). *Countering Online Hate Speech*. Paris: UNESCO Publishing.

Goodman, E. and Cherubini, F. (2013). *Online Comment Moderation: Emerging Best Practices*. Germany: Darmstadt, The World Association of Newspapers WAN-IFRA. Available from <http://www. wan-ifra. org/reports/2013/10/04/online-comment-moderation-emerging-best-practices>.

Hechter, M. and Opp, K.-D. (2001). *Social Norms*. New York: Russell Sage Foundation.

Heckathorn, D. D. (1988). Collective sanctions and the creation of prisoner's dilemma norms. *American Journal of Sociology*, **94**, 535–562.

Henson, B., Reyns, B. W. and Fisher, B. S. (2013). Fear of crime online? Examining the effect of risk, previous victimization, and exposure on fear of online interpersonal victimization. *Journal of Contemporary Criminal Justice*, **29**.

Hinduja, S. and Patchin, J. W. (2007). Offline consequences of online victimization: school violence and delinquency. *Journal of School Violence*, **6**, 89–112.

Ivarsaten, E., Blinder, S. and Ford, R. (2010). The anti-racism norm in western European immigration politics: why we need to consider it and how to measure it. *Journal of Elections, Public Opinion and Parties*, **20**, 421–445.

Judd, C. M., Westfall, J. and Kenny, D. A. (2017). Experiments with more than one random factor: designs, analytic models, and statistical power. *Annual Review of Psychology*, **68**, 601–625.

Keizer, K., Lindenberg, S. and Steg, L. (2008). The spreading of disorder. *Science*, **322**, 1681–1685.

Kennedy, M. A. and Taylor, M. A. (2010). Online harassment and victimization of college students. *Justice Policy Journal*, **7**, 1–21.

Keuschnigg, M. and Wolbring, T. (2015). Disorder, social capital, and norm violation: three field experiments on the broken windows thesis. *Rationality and Society*, **27**, 96–126.

Kraut, R. E. *et al.* (2012). *Building successful online communities: Evidence-based social design*. Cambridge, MA and London: MIT Press.

Krippendorf, K. (2004). Reliability in content analysis. *Human Communication Research*, **30**, 411–433.

Krupka, E. and Weber, R. A. (2009). The focusing and informational effects of norms on pro-social behavior. *Journal of Economic Psychology*, **30**, 307–320.

Krupka, E. L. and Weber, R. A. (2013). Identifying social norms using coordination games: why does dictator game sharing vary? *Journal of the European Economic Association*, **11**, 495–524.

Mantilla, K. (2013). Gendertrolling: misogyny adapts to new media. *Feminist Studies*, **39**, 563–570.

Matias, J. N. (2016). *Posting Rules in Online Discussions Prevents Problems & Increases Participation*. Available from: https://civilservant.io/moderation_experiment_r_science_rule_posting.html.

Munger, K. (2016). Tweetment effects on the tweeted: experimentally reducing racist harassment. *Political Behavior*, **39**, 629–649.

Paluck, E. L. and Green, D. P. (2009). Prejudice reduction: what works? A review and assessment of research and practice. *Annual Review of Psychology*, **60**, 339–367.

Pettigrew, T. F. (1958). Personality and sociocultural factors in intergroup attitudes: a cross-national comparison. *Journal of Conflict Resolution*, **2**.

Pettigrew, T. F. (1991). Normative theory in intergroup relations: explaining both harmony and conflict. *Psychology and Developing Societies*, **3**, 3–16.

Rauhut, H. and Winter, F. (2012). On the validity of laboratory research in the political and social sciences: the example of crime and punishment. In *Experimental Political Science*. London: Palgrave Macmillan, pp. 209–232.

Ross, B. *et al.* (2016). Measuring the reliability of hate speech annotations: the case of the european refugee crisis. In Bei_wenger, M., Wojatzki, M. and Zesch, T. (Eds.), *Proceedings of NLP 4 CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, Vol. **17**. Bochum, pp. 6–9.

Rost, K., Stahel, L. and Frey, B. S. (2016). Digital social norm enforcement: online firestorms in social media. *PLoS One*, **11**, e0155923.

Schieb, C. and Preuss, M. (2016). Governing hate speech by means of counterspeech on Facebook. In *66th ICA Annual Conference, At Fukuoka, Japan*, pp. 1–23.

Shapiro, J. R. and Neuberg, S. L. (2008). When do the stigmatized stigmatize? The ironic effects of being accountable to (perceived) majority group prejudice- expression norms. *Journal of Personality and Social Psychology*, **95**, 877.

Shadish, W. R., Cook, T. D. and Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.

Sherif, M. and Sherif, C. W. (1953). *Groups in Harmony and Tension; an Integration of Studies of Intergroup Relations*. New York: Harper and Brothers.

Snijders, T. A. (2011). *Multilevel Analysis*. New York: Springer.

Stangor, C., Sechrist, G. B. and Jost, J. T. (2001). Changing racial beliefs by providing consensus information. *Personality and Social Psychology Bulletin*, **27**, 486–496.

Tankard, M. E. and Paluck, E. L. (2016). Norm perception as a vehicle for social change. *Social Issues and Policy Review*, **10**, 181–211.

Voss, T. (2001) Game-theoretical perspectives on the emergence of social norms. In Hechter, M. and Opp, K.-D (Eds.), *Social Norms*. New York: Rusell Sage Foundation.

Winter, F., Rauhut, H. and Miller, L. (2017). *Dynamic Bargaining and Normative Conflict*. Max-Planck-Institute for Research on Collective Goods Working Paper. Bonn, Germany.

Zitek, E. M. and Hebl, M. R. (2007). The role of social norm clarity in the influenced expression of prejudice over time. *Journal of Experimental Social Psychology*, **43**, 867–876.

Amalia Álvarez-Benjumea is a PhD-Candidate in the Max Planck Research Group on "Mechanisms of Normative Change" at the Max Planck Institute for Research on Collective Goods, Bonn, Germany, and the University of Cologne. Her research interests include methods of online experiments, the effect of social information on normative behaviour, perception of norms, and opinion dynamics under normative and social influence.

Fabian Winter is head of the Max Planck Research Group on "Mechanisms of Normative Change" at the Max Planck Institute for Research on Collective Goods, Bonn, Germany. Current research interests comprise social norms, normative conflicts, quantitative and experimental methods, modelling social processes and big data analysis. His work has been published among others in *Social Forces, Economics Letters, Social Science Research* and *Mathematical Sociology*.