



Normative change and culture
of hate: An experiment on
online environments

Amalia Álvarez
Fabian Winter





Normative change and culture of hate: An experiment on online environments

Amalia Álvarez / Fabian Winter

February 2018

This article has been accepted for publication in *European Sociological Review* Published by Oxford University Press. <https://doi.org/10.1093/esr/jcy005>

Normative change and culture of hate: An experiment in online environments

Amalia Álvarez-Benjumea ^{*1} and Fabian Winter¹

¹Max Planck Institute for Research on Collective Goods

First draft: June 2017

This Draft: February 2018

This article has been accepted for publication in European Sociological Review Published by Oxford University Press. <https://doi.org/10.1093/esr/jcy005>.

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: 1) what do others normally do, i.e., descriptive norms; and 2) 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 normative behavior 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.

Keywords: *online experiments, social norms, hate speech, social influence, pluralistic societies*

^{*}corresponding author: alvarezbenjumea@coll.mpg.de

1 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 community¹, or other minority groups (Kennedy & Taylor, 2010; Mantilla, 2013). This behavior 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 behavior, such as online harassment and trolling (Binns, 2012), since people who engage in these types of behavior 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 behaviors. We define hate speech as hostile behavior and “antagonism towards people” (Gagliardone, Gal, Alves, & Martinez, 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 & Patchin, 2007) and push people into withdrawing from the public debate, therefore harming free speech (Henson, Reyns, & 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 speech often involve favoring counter-speaking (flagging, reporting, etc.) or censoring the hate content (Citron & Norton, 2011; Goodman & Cherubini, 2013). The theoretical and policy-making 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.

¹Lesbian, Gay, Bisexual and Transgender.

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

³UN 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>).

⁴Concerns 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 & 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).

⁵Different interventions have been discussed in the literature. Goodman and Cherubini (2013), for example, 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 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).

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 behavior is common brings about more antisocial behavior, such as littering, stealing, or jaywalking (Cialdini, Reno, & Kallgren, 1990; Keizer, Lindenberg, & Steg, 2008; Keuschnigg & Wolbring, 2015). A similar process has been found in online contexts, where prior troll comments affect the likelihood of subsequent trolling (Cheng, Danescu-Niculescu-Mizil, Leskovec, & Bernstein, 2017). This cascading dynamic is linked to a process of spreading norm violations: people learn from each other which kind of behavior is approved and which behavior 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 behavior as socially accepted. The opposite should hold true: reducing the social acceptability of hateful behavior 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 (e.g., Pettigrew, 1991; Paluck & Green, 2009). Highlighting a majority norm, or a perceived consensus against the expression of prejudice, reduces people’s willingness to express prejudice (Crandall, Eshleman, & O’Brien, 2002; C. Crandall & Stangor, 2005; Stangor, Sechrist, & Jost, 2001; Shapiro & 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. 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 behavior of interest normally occurs, which increases ecological and external validity (Shadish, Cook, & Campbell, 2002; Rauhut & Winter, 2012).

2 Theory and Hypotheses

Social norms are shared rules that provide the standard of behavior within a wide range of settings (Elster, 1989; Coleman, 1990a, 1990b; Hechter & Opp, 2001; Bicchieri, 2005), and the behavior 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, List, Malmendier, & Rao, 2016), to avoid rejection (Cialdini & Goldstein, 2004), or to avoid further sanctioning (Hechter & Opp, 2001). Norms are usually not clearly determined and individuals rely on their subjective perceptions of norms (Tankard & Paluck, 2016); people use the behavior 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 (Bursztyn, Egorov, & Fiorin, 2017; Cialdini & Goldstein, 2004; Cialdini & Trost, 1998). Likewise, prejudice is subject to similar normative influence.⁷ There

⁶Social 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 & McNally, 2016, p. 2).

⁷Allport (1979) and Sherif and Sherif (1953) wrote seminal texts arguing that the majority of prejudiced attitudes arose from

is evidence of a norm against public expression of hate in Europe (Ivarsflaten, Blinder, & Ford, 2010; Blinder, Ford, & Ivarsflaten, 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 behavior 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 et al. (2017) argue that the 2016 election results in the USA causally increased individuals' perception of the social acceptability of anti-immigration views and "their willingness to publicly express them".

Observing the behavior 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, Crandall, Brigham, & Vaughn, 1994, p. 362), showing consensus information over negative stereotypes (Stangor et al., 2001), or hearing others endorse an anti-prejudice norm (Zitek & 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 & Trost, 1998).

2.1 Conveying Information about Appropriateness

We have argued that individuals infer acceptability from the context using the behavior of other actors within it as a source of normative information. Previous literature has identified two sources of information: 1) what do others normally do, and 2) 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 (e.g., Cialdini et al., 1990; Cialdini & Goldstein, 2004; Bicchieri & Xiao, 2009). Descriptive norms act as coordination devices of "normal behavior" (e.g., Bicchieri, 2005; Krupka & Weber, 2013), whereas injunctive norms act as an "oughtness rule" (e.g., Hechter & Opp, 2001). Situational triggering cues can also increase the saliency of normative information and reduce ambiguity about the appropriateness of a certain type of behavior. Actions that stand out, such as observing someone punishing, draw attention to the existing norm. The implications for online behavior are straightforward. While observing prevalent online behavior illustrates the descriptive norm, observing responses to those behaviors 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 positive relationship between the subject's action and what she observes others have done, thus:

conformity to social normative expectations. In their work, 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.

⁸Rejection of public expression of prejudice has been generally increasing in the last 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; Duckitt, 1992; Dovidio & Gaertner, 1986)

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 behavior 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, & Frey, 2016), also against hate speech (Schieb & Preuss, 2016). If individuals need to see the consequences of behavior 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.*

3 Experiment

3.1 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 analyzed 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 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 time line is described in [Appendix A](#).¹¹

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

¹⁰Pictures 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](#)).

¹¹The survey was conducted in April 2016. The pre-experimental collection of the comments was carried out on June 2016. Finally, the experiment was conducted on 4 and 5 August 2016.

3.2 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.

Treatment	Summary of forum content
<i>Baseline</i>	6 comments: 2 friendly, 2 neutral, and 2 hostile
<i>Censored</i>	4 comments: 2 friendly, and 2 neutral
<i>Extremely censored</i>	3 comments: all friendly
<i>Counter-speaking</i>	6 comments: 1 friendly, 1 neutral, and 2 sanctions

Table 1: Summary of the content of the online forum in the different treatments

The exact comments and the order of appearance in the discussion were automatically selected from a database for each participant in the experiment.¹² Table 1 summarizes the number of comments in the different experimental conditions.

3.3 Data Collection

The experiment was conducted entirely online and participants were recruited from a crowdsourcing Internet marketplace.¹³ The experiment was conducted entirely in German and the sample was restricted to German

¹²As 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.

¹³We 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. The disad-

residents. Although we did not directly ask participants for their demographic characteristics, the subjects were selected from a population with the characteristics depicted in Table 2. The sample is obviously more diverse than the traditional convenience sample of students.

Gender	
Women	55%
Age	
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%

Table 2: Sociodemographic Characteristics of the Population from where Subjects Were Recruited

Participants were compensated with a fixed amount of three 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.

Treatment	Subjects	Comments
Baseline	47	375
Counter-speaking	45	373
Censored	46	377
Extremely censored	42	344
Total	180	1469

Table 3: Number of participants and valid comments per treatment

A total of 180 participants were recruited to take part in the forum. Participants spent an average of ten vantage was that using online workers in the experiment could also raise concerns about the external validity of the experiment. Findings in this experiment may not generalize to everyone in every context, but we argue that they generalize to users in online forums because, despite being paid, participants remained anonymous and could abandon the experiment at any moment, which makes their comments voluntary. Since the comments were voluntary, we believe that their motivation to express hate should not differ from motivations of the general population of online forum commenters and, if they do, these differences are constant among treatments. Furthermore, using online workers is, if anything, underestimates the prevalence of hate speech. Since we are not interested in prevalence estimates, but in the effectiveness of our treatments, we do not consider this a problem.

minutes in the forum. We collected a total of 1585 comments, of which 116 were invalid.¹⁴ 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.

3.4 Measurement of Variables and Operationalization

We evaluated the comments in two ways: we assigned them a score, using a 9-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 (2016), 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 1029: “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 speech have been found not to be very reliable (Ross et al., 2016). Inter-rater reliability of our scale is relatively high (Krippendorff’s $\alpha = 0.69$).¹⁵ 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: 1) contains negative stereotypes, 2) uses racial slurs, 3) Contains words that are insulting, belittling, or diminishing, 4) calls for violence, threat, or discrimination, 5) uses sexual slurs, and 6) sexual orientation/gender used to ridicule or

¹⁴The 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. Participants were asked to comment on 9 different threads. We ran the analysis excluding the comments of those who failed to leave 9 comments ($N=6$), and the results did not change.

¹⁵We computed different measures of inter-rater agreement and reliability such as intra-class correlation ($ICC=0.704$). We chose Krippendorff’s α to assess inter-rater reliability. This measurement is commonly used by researchers in content analysis (Krippendorff, 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 ($\alpha = 0.71$) and the lowest in LGBT ($\alpha = 0.58$).

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 (2016).¹⁶ Comments were labeled 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 labeled as hateful also increases.¹⁷ 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 A](#).

4 Data and Results

4.1 Average Levels of Hate Speech

We begin this section by analyzing 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 (blue line) for all topics except poverty.

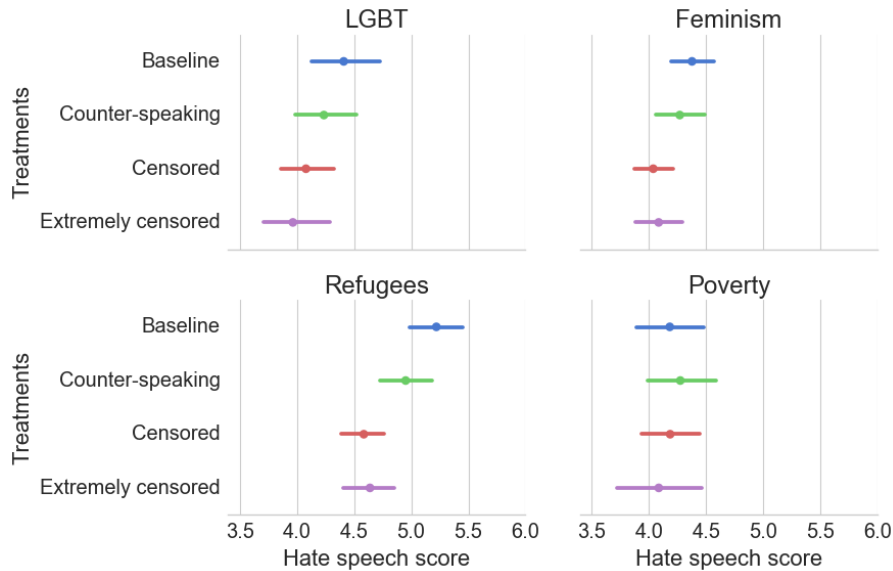


Figure 1: Treatment differences in mean hate speech score across topics (Obs=1469). Error bars at the 95% confidence interval.

¹⁶ “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).

¹⁷ The predicted probabilities of a comment being classified as hateful by the three raters increases from 0.054 at a score of 6 to 0.98 at a score of 9. Only the 5.5% of comments containing an item from the list has an average score of 5 or lower.

We estimated a random intercept multilevel regression model (Judd, Westfall, & 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).¹⁸ 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 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:

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

$$Y_{ijk} = \beta_0 + \beta_1 \text{Treatment}_{ij} + \beta_2 \text{Topic}_{ijk} + u_i + v_j + \epsilon_{ijk} \quad (2)$$

$$Y_{ijk} = \beta_0 + \beta_1 \text{Topic}_{ijk} + \beta_2 (\text{Treatment}_{ij} * \text{Topic}_{ijk}) + u_i + v_j + \epsilon_{ijk} \quad (3)$$

$$\text{where } u_i \sim N(0, \sigma_u) \text{ and } v_j \sim N(0, \sigma_j) \quad (4)$$

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/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.

¹⁸Subjects were asked to leave 9 comments in 9 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 ANOVA. Both random levels are significant. The magnitude of the intra-class correlation (ICC) estimate, i.e., variance accounted for by between-subjects differences, suggests that variability between subjects is very high and should be taken into account in all analysis

Table 4: Results from Multilevel Random Models of Hate Speech Score

	Model 1	Model 2	Model 3
<i>Main effects</i>			
Constant	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)*	
LGTB		-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)
<i>Interaction effects</i>			
Poverty*Counter-speaking			0.07 (0.24)
LGTB*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)
LGTB*Censored			-0.33 (0.20)
Refugees*Censored			-0.64 (0.19)**
Feminism*Censored			-0.33 (0.18)
Poverty*Extremely			-0.12 (0.25)
LGTB*Extremely			-0.45 (0.21)*
Refugees*Extremely			-0.60 (0.19)**
Feminism*Extremely			-0.28 (0.19)
Groups:Subjects	180	180	180
Var:Subjects	9	9	9
Groups:Subjects	0.44	0.44	0.44
Var:Subjects	0.15	0.11	0.11
Residual Variance	0.90	0.90	0.90
ICC: Subjects	0.30	0.30	0.30
ICC: Pictures	0.10	0.07	0.07
AIC	4345.59	4347.50	4371.48
BIC	4382.64	4400.42	4472.04
Log Likelihood	-2165.80	-2163.75	-2166.74
Obs	1469	1469	1469

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. Significance levels: *** $p < 0.000$, ** $p < 0.01$, * $p < 0.05$, † $p < 0.1$, for a two-sided test

4.2 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 baseline a Kolmogorov-Smirnov test of equality of the distributions yields $P < 0.001$).

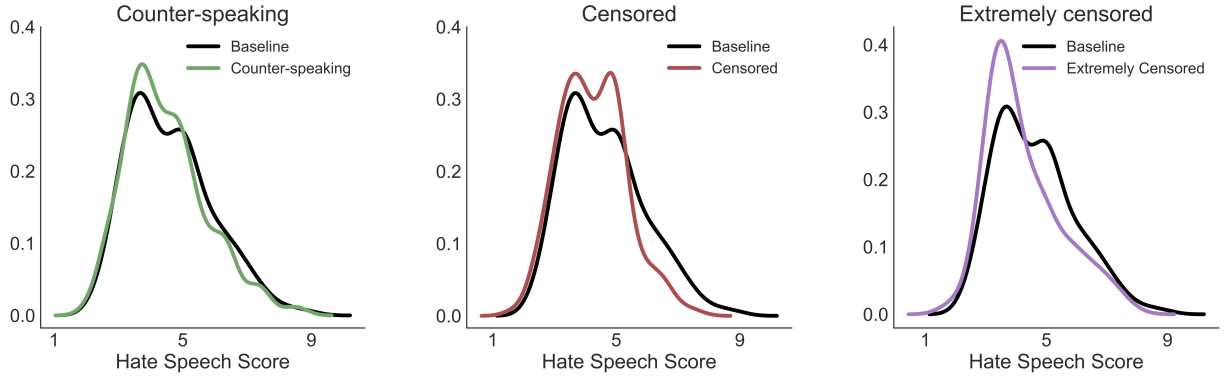


Figure 2: Density estimates of average hate speech score in the counter-speaking treatment (green), the censored condition (red), and the extremely censored condition (purple), compared to baseline treatment (black). Size of bins selected using Freedman–Diaconis rule. All treatments are compared to the baseline group ($N=1469$). The graph depicts the 1 to 9 score scale: scores on the left correspond to friendly speech, while scores on the right correspond to more hostile language

Next, we analyze 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%), compared to 4 comments in the censored condition (0.99%). The reduction is significant [$\chi^2(1, 814) = 14.94, P < 0.001$].¹⁹ There are 18 hostile comments in the counter-speaking treatment (4.44%) and 15 in the extremely censored treatment (4.10%). None of them significantly reduces extremely hostile comments. Similar results are obtained using a quantile regression. We computed the treatment effect in the .75, .90, .95, and .99 quantiles of the hate score distribution. The censored condition significantly reduces hate in the .75, .90, .95, and .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 scores points in the 0.99 quantile compared to the baseline (see [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 1 or 2 in total. The maximum number of hostile comments left by one unique participant is 7. [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 [$\chi^2(1, 721) = 7.63, P < 0.01$]; 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 analyze the distribution of hostile comments from the participants’ perspective (see [Appendix B](#)).

Distribution of Hostile comments per participant

Treatments	n=0	n=1	n=2	n=3	n=4	n=5	n=6	n=7
Base	36	6	0	4	0	0	1	0
Counter-speaking	34	7	3	0	0	1	0	0
Censored	43	2	1	0	0	0	0	0
Extremely Censored	34	6	1	0	0	0	0	1

Table 5: Distribution of number of hostile comments per participant

¹⁹Our findings are robust to the different inference methods as displayed in [Table 9](#) in [Appendix B](#). The effect of the censored treatment is robust to the removal of influential individuals. These analyses are available upon request

4.3 Analysis of the Norm Violations

In addition to the hate score, we analyzed comments that were classified as a norm violation according to our hate speech indicators, that is, comments that are regarded as uncivil.²⁰ 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).

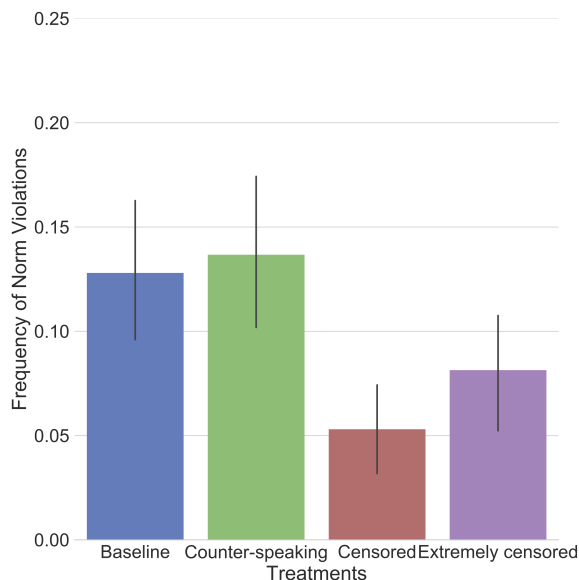


Figure 3: Proportion of comments that were labeled as hate speech across treatments (N=1469). bars at 95% CI

We tested for differences in the frequency of hate comments among the different treatments, using a multilevel logistic model with a random effect for participants (Table 6).²¹

²⁰The agreement between the raters is low (Krippendorff's $\alpha = 0.40$), (Ross et al., 2016) found that inter-rater agreement for binary classification of tweets as hateful or not hateful is very low

²¹Our results are also robust to different testing methods (See Table 10 in Appendix B), and to the removal of the most influential individuals (analyses available upon request).

Table 6: Results from a Multilevel Logistic Model of the Probability of a Norm Violation

Model (1)	
Main effects	
Constant	-2.53 (0.30)***
Counter-speaking	0.19 (0.38)
Censored	-1.00 (0.42)*
Extremely censored	-0.50 (0.41)
Random Pars	
Groups: Subjects	180
Var: Subjects	1.57
Groups: Pictures	9
Var: Pictures	0.05
AIC	892.46
BIC	924.22
Obs.	1469

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. Significance levels: *** $p < 0.000$, ** $p < 0.01$, * $p < 0.05$, † $p < 0.1$, for a two-sided test

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 of the number of norm violations compared to the baseline condition. Again, we find support for the 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.

Total number of Uncivil Comments per Participant								
Treatments	n=0	n=1	n=2	n=3	n=4	n=5	n=6	n=7
Baseline	26	10	3	3	3	1	1	0
Counter-speaking	20	10	9	4	1	0	0	1
Censored	33	7	5	1	0	0	0	0
Extremely Censored	23	16	0	1	1	1	0	0

Table 7: Distribution of number of norm violations per participant

4.4 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.

Second, 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 labor 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 paper 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 does 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. This data would be observational, with the endogeneity problem that goes along with it. By contrast, our data is collected in a controlled experimental environment, and therefore allows for a proper identification of the treatment effect.

5 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 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 (H 1a), 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 than in the censored condition, which hints at a polarization of opinions. Our intuition is that this result might indicate either a reactance effect ²² or an increase in normative ambiguity.

²²Reactance appears when an individual facing a persuasive message reacts by engaging in the proscribed behavior (Burgoon,

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 allows for a 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 behavior, 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 favors them (Xiao & Bicchieri, 2010; Bicchieri & Chavez, 2013; Winter, Rauhut, & 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 behavior. Our results suggest that normative behavior 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 behavior (Heckathorn, 1988; Coleman, 1990b; 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 & Weber, 2013). 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 prevent 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 behavior act as a situational normative cue in online environments. Second, our findings point to a larger effect of descriptive norms — defined as frequent behavior — 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 tradeoff between censoring and free speech; hence, our paper does not represent a position on whether censoring hate content is necessarily socially beneficial. We consider that a social norm intervention (e.g., Tankard & Paluck, 2016) is a good approach to address online hate speech, whose presence is not necessarily considered unlawful, but often regarded as undesired.

A Appendix A

A.1 Keywords Used in Pictures Search

The images were obtained from Twitter and Google images in March 2016. We used different tags or keywords to search for images that were twitted using them. We used both German and English terms that are often used on German social media.

Sharia, Multiculturalism, Terrorism, Religion, Transgender, Gay, Homosexuality, Sexism, Discrimination, Refugees, *aufschrei*, Sexism, Immigration, homosexuality, *Einwanderung*, Diversity, Queer, Begging, Atheism, islamization, Religion, *#tolerist*

A.2 Online Survey

The following questionnaire was answered by 90 participants during the pre-experimental phase.

Survey (*In German in the original*)

1. How controversial do you think the picture above is? (on a scale from 1 to...)
2. Would you say this picture represent more a positive or negative view of the issue depicted in the image?
3. Which topic would you say the picture above represents best?
(Choose only one)
 - Refugees
 - Immigration
 - Gendermainstreaming
 - Transgender
 - Genderbender
 - *#aufschrei*
 - Aggressive behavior
 - Romany
 - Begging
 - Grexit
 - Eurocrisis

- Zionism
- Judaism

A.3 Instructions

Einleitung

Vielen Dank für Ihre Teilnahme.

Wir werden Ihnen im Folgenden eine Reihe von Bildern zeigen und Sie nach Ihrer Meinung zu diesen Bildern befragen. Bitte lesen Sie die nachfolgenden Anweisungen sorgfältig durch, bevor Sie mit der Bearbeitung des Fragebogens beginnen.

Ihre Teilnahme ist für uns sehr wichtig. Jegliche Information, die Sie uns während der Beantwortung des Fragebogens geben, wird streng vertraulich behandelt und wird alleinig zum Zweck unserer Studie verwendet. Ihre Angaben werden gemäß den in Deutschland einschlägigen Richtlinien zum Datenschutz gespeichert.

Im Folgenden wird Ihnen ein zufällig ausgewählter Benutzername zugeordnet, unter welchem Ihre Daten gespeichert und angezeigt werden.

Zum Ende dieses Fragebogens werden Sie einen Identifizierungscode erhalten, mit dem Sie bei clickworker.com Ihre Auszahlung anweisen können.

Weiter

Figure 4: Introduction Page of the Experiment

Teil 2

Im Folgenden wird Ihnen eine Reihe von Bildern gezeigt, an die sich jeweils eine Diskussion anschließt. Ihre Aufgabe ist es, ebenfalls einen Kommentar in dieser Diskussion zu verfassen.

Ihnen wird für die Dauer dieser Aufgabe ein Nutzernamen und ein Nutzersymbol zugeordnet, die zur Ihrer Identifikation während der Diskussion dienen. Andere Teilnehmer können so auf Ihre Kommentare reagieren. Sowohl der Nutzernamen als auch das Symbol können allerdings nicht mit Ihrer realen Identität in Zusammenhang gebracht werden, so dass Sie anonym bleiben.

Ihr Kommentar sollte aus mindestens zwei bis drei Sätzen bestehen. Diese Sätze sollten aussagekräftig sein und sich auf die Bild bzw. die Diskussion beziehen.



User1: Ich weiß, dass viele Leute Graffiti mögen und sie sogar als Kunst betrachten. Allerdings kann ich Graffiti gar nicht leiden und denke, dass sie die Städte verschandeln.

User2: Das denke ich auch. Die Stadtverwaltung sollte sowas entfernen lassen.

User3: Einige dieser "Graffiti" werden noch in Museen zu sehen sein. Sie repräsentieren die wirkliche moderne Kunst. Das sollte jeder verstehen.

Bitte hinterlassen Sie Ihren Kommentar.

Ein aussagekräftiger Kommentar zu dem oben gezeigten Bild wäre zum Beispiel:

„Ich weiß, dass einige Leute das schön finden, aber für mich ist es bloße Schmiererei! Es verschandelt die Städte. Die Politik sollte endlich etwas dagegen unternehmen.“

Dies hier wäre auch in Ordnung:

„Ich verstehe die Meinung im ersten Kommentar, aber ich stimme dem nicht zu. Ich finde das schön. Es gehört doch heute einfach mit dazu. Ausserdem sollten junge Leute auch ihre Freiräume haben um sich auszuprobieren.“

Der folgende Kommentar hingegen würde nicht als zulässig eingestuft werden:

„Der schnelle braune Fuchs springt über den faulen Hund. Der schnelle braune Fuchs springt über den faulen Hund. Der schnelle braune Fuchs springt über den faulen Hund.“

Ebenso wäre folgender Kommentar kein zulässiger Kommentar:

„Verlassener Ort!“

Jede Seite wird nur einmal angezeigt. Nachdem Sie Ihren Kommentar abgegeben haben, können Sie zur nächsten Seite wechseln. Sie können jedoch nicht zurückgehen oder vorherige Kommentare bearbeiten.

Bitte drücken Sie "Weiter", wenn Sie bereit sind mit dem Fragebogen zu beginnen. Vielen Danke!

Weiter

Figure 5: Instructions of the Experiment

Introduction (In English)

Thank you for your participation.

We will show you a series of pictures and ask you comment on them. Please read the following instructions carefully before you begin the task. Your participation is very important to us. Any information you provide to us during the task will be strictly confidential and will be used solely for the purpose of our study. Your data will be stored in accordance with the relevant data protection guidelines in Germany.

You will be assigned a random user name, and your input will be stored and displayed under this username. At the end, you will be given an identification code, which will allow you to claim your payment at clickworker.com

Instructions (In English)

You will see a series of pictures with a discussion below. Your task is to join the discussion on the topic(s) depicted in the picture(s). Please write at least two to three sentences per discussion. These sentences should be meaningful and relate to the picture/discussion.

A valid comment on the discussion above would be:

“I know that some people like them and even consider them to be art. However, I really dislike graffiti or “street art” and some call it. I think it impoverishes the way a city looks”

“I do understand the opinion in comment 1, although I pretty much disagree. Most of the places that are now covered by graffiti were previously abandoned and looked very dirty and ugly already”

The following comment would not count as valid:

“The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog.”

The following is also not a sufficient comment:

“Abandoned place”

Each page will be shown just once. Once you have finished with your comment you can go to the next page, but you cannot go back or edit previous comments.

Please press the “start” button once you are ready to start the survey.

Thanks!

A.4 Screenshots of the Experiment

Bitte beteiligen Sie sich nun mit einem Kommentar an der Diskussion.



**Nicely**

Das Bild könnte aus Griechenland stammen und einen Aufstand der Flüchtlinge zeigen, die mit den Bedingungen unter denen sie leben müssen, nicht einverstanden sind

**Lorely**

warum müssen die flüchtlinge alles zerstören, nur weil es nicht so läuft wie sie es verlangen. denen möchte ich abends nicht begegnen. das sind leute, die in köln am silvesterabend frauen sexual bedrängt haben.

**Strohblume**

Migranten versuchen mit Gewalt einen Grenzzaun einzureißen. Nur konsequente staatliche Gewalt kann hier helfen. Auch sollten die Gewalttäter mit herben Konsequenzen in Form von Abschiebungen belangt werden.

**Kaktusstachel**


Flüchtlinge die versuchen einen Sperrzaun umzureißen, damit sie ihre Flucht fortsetzen können

**Nicely**

Jeder sollte die Möglichkeit bekommen, sich ein sicheres Zuhause zu suchen. Grenzen sollten nicht geschlossen werden.

**usertrench**

Ich möchte nicht in der Haut der verzweifelten Flüchtlinge stecken. Auf der einen Seite von Krieg und Tod vertrieben auf der sicheren Seite nicht willkommen, sehr menschenunwürdig!

**userreceived**

bitte ausfüllen

weiter

Figure 6: Baseline treatment

Bitte beteiligen Sie sich nun mit einem Kommentar an der Diskussion.



userrench

Zu dem Bild kann ich nicht all zu viel sagen, ich sehe eventuell Homosexuelle Menschen auf einem Marsch? oder einer Protestaktion.



stipulateuser

Homosexuelle müssen heutzutage leider immer noch für ihr Recht und ihre Anerkennung auf die Straße gehen. Ich hoffe, das wird sich in naher Zukunft ändern.



Freiluft

Kundgebungen solcherart kann ich nur befürworten. Jede öffentliche Demonstration sorgt für eine breitere Akzeptanz und sorgt dafür, dass sich mehr Menschen outen.



userdext

Die Person hat das Wort Freedom auf den Bein stehen und generell verbinde ich das Auftreten der Person als sehr offen gegenüber anderen. Dies scheint eine Art Schwulenparade oder Schwulendemo zu sein.



Nicely

bitte ausfüllen

weiter

Figure 7: Censored treatment

Bitte beteiligen Sie sich nun mit einem Kommentar an der Diskussion.



Silberstreif

Die westliche Dame scheint neugieriger zu sein als die muslimische Frau. Da ich in einer Großstadt lebe, bekomme ich dieses Gegensätzlichkeit oft mit, aber stören tut es mich nicht.



Pretty

Was für ein kultureller Unterschied. Ich frage mich allerdings wie man damit einen Apfel isst.



Goldroute

Ich denke, die blonde Frau ist verunsichert, da sie nicht die Mimik der verhüllten Frau deuten kann. Zudem scheint ihr der Anblick fremd zu sein. Ich persönlich stehe dem Bild neutral gegenüber



userrcutel

bitte ausfüllen


weiter


Figure 8: Extremely Censored treatment


Bitte beteiligen Sie sich nun mit einem Kommentar an der Diskussion.




**usericomly**
Frauen und Transsexuelle in der Armee sind absolute Ausnahmen. Ich habe da nichts gegen, wobei ich Militarismus grundsätzlich ablehne.

**userhandlers**
Es ist schön zu sehen, dass selbst in so konservativen Bereichen wie der Verteidigung Toleranz einkehrt

**Kekskuchenfan**
Ein Mann in Frauenpose ist lächerlich.

**Mohnblume**
Es ist einfach nur ungewohnt und entspricht nicht unserem Weltbild, aber das macht es nicht lächerlich.

**userDj**
Ich bin gegen Transgender im Berufsleben. Was die Menschen privat machen, ist ihre Sache, auch wenn ich auch das schon abstoßend finde.

**Silberstreif**
Es hat nichts mit "gegen Transgender im Berufsleben" zu tun, das hat generell was mit Akzeptanz zu tun was Sie hier beschrieben. Jeder Mensch darf sich frei entfalten, das gilt auch für das Berufsleben, ob Mensch im flaschen Körper oder homosexueller.

**Pfefferminze**
bitte ausfüllen




Figure 9: Counter-speaking treatment

A.5 Examples of Comments

Hate comments

Comment 1261 (User 171): "Flüchtlingskrise. Die können weiter wandern weg von europa. das sind nicht nur kriegsflüchtlinge, zu 90 prozent sind es nur sozialschmarotzer die hier machen können was die wollen." (Hate score: 8.66)

(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.)

Comment 1029 (User 112): “Schwule Kerle sind das Letzte, was ich dulden würde. Schon gar nicht in der Öffentlichkeit.” (Hate score:9)

(Gay guys are the last thing I would tolerate, especially not in public.)

Neutral comments

Comment 892 (User 179): “Generell bin ich dagegen sich in der Öffentlichkeit wild zu küssen. Aber gegen einen Kuss habe ich nichts.” (Hate score: 5)

(In general I am against passionate kissing in public. But I have nothing against a kiss.)

Friendly comments

Comment 1572 (User 71): “Das ist wirklich eine wunderschöne Szene. Es sieht nach einer absolut glücken Familie aus. Wahrscheinlich sind sie glücklicher als so manches hetero Paar.”(Score of 1,66)

(This is really a wonderful scene. It looks like an absolutely happy family. They are probably happier than many heterosexual couples.)

Comment 216 (User 63): “Super Daumen hoch für diese Leute die den Mumm haben sich der Ignoranz zu stellen.” (Score of 2)

(Super thumbs up for these people, who have the guts to face up to this ignorance.)

A.6 Timeline of the Experiment

The process is divided into three phases: collection of materials, treatment design, and collection of comments and analysis. All pre-experimental sessions were conducted using workers recruited from Clickworker. During the collection of materials (March to April, 2016), we collected images using the keywords in A1. Pictures were collected from Google images and Twitter. The pictures were classified into different categories (see categories in question 3 of the online survey in A2). In April 2016, we ran an online survey (N=90) with a sample of German speakers from Clickworker and selected the pictures and topics rated as “more controversial”. In June 2016 we made the forum with the 9 selected pictures available online to workers. The comments collected in this first pre-experimental session were rated by three independent raters using the hate score (1 to 9 scale), and classified into 3 categories: friendly, neutral, and hostile. In a second pre-experimental session using workers, we collected replies to hostile comments, which could be used as verbal sanctions to construct the counter-speaking condition. We used them to construct the counter-comments condition. All experimental conditions are constructed using previous comments. The exact comments and the order of appearance in the discussion were automatically selected each time a participant joined the experiment, e.g. when a participant was allocated to the extremely censored condition she could see 3 randomly chosen friendly comments.

Finally, we conducted the experiment during the 4-5th August 2016, we made our forum available online and distributed the link to the participants in the experiment (N=180) via Clickworker. The raters then classified the collected comments during a period of 2 weeks. The score and items were finally analyzed.

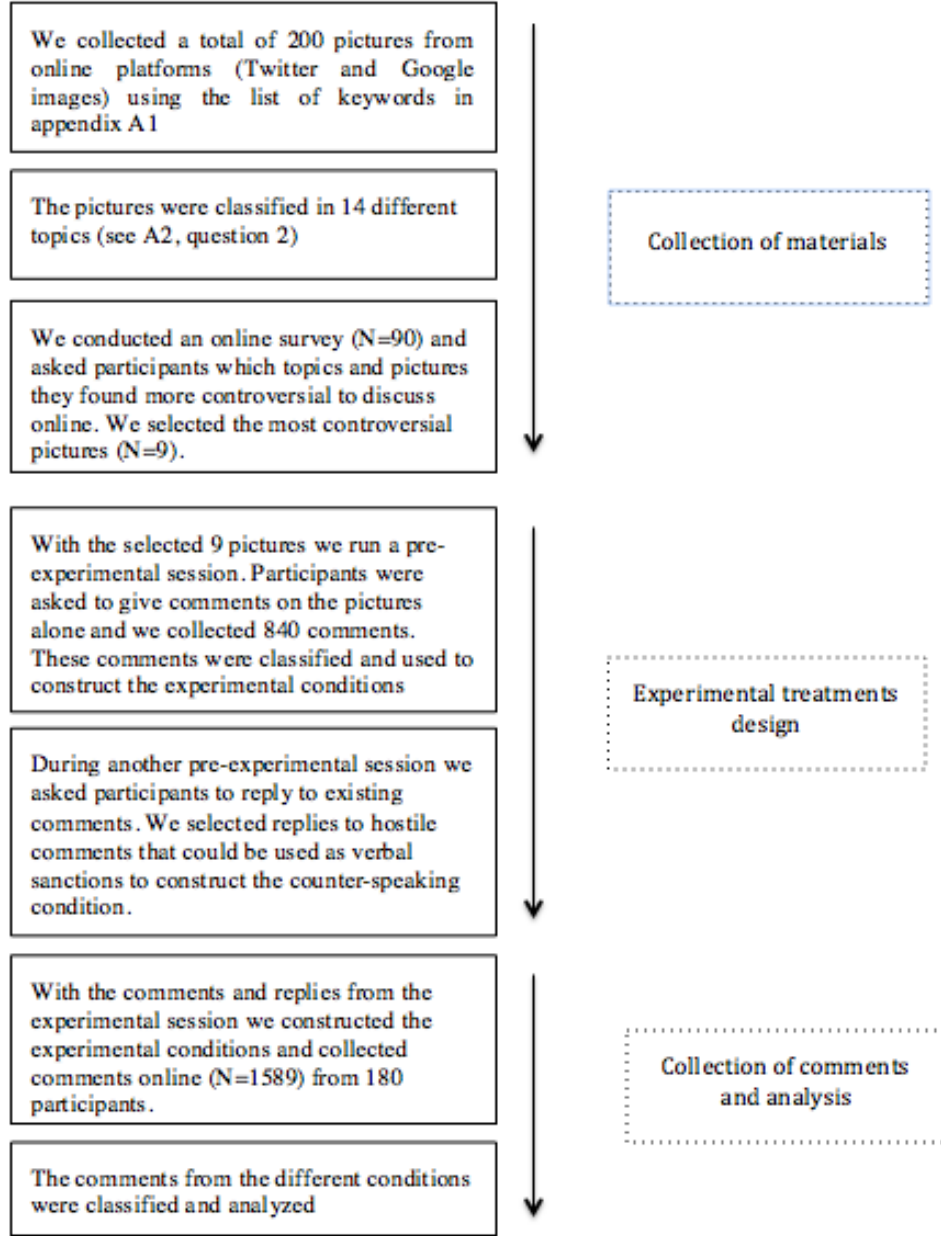


Figure 10: Timeline of the Experiment

B APPENDIX B

B.1 Analyses of Extremely Hateful Comments

We estimated linear conditional quantile regressions at the 0.75, 0.90, 0.95 and 0.99 quantiles of the hate score distribution. The table shows the treatment effects estimates for each of the quantiles or quantile coefficients.

This gives a more complete description of how the effect of the treatments work at the higher quantiles, which correspond to the most hostile comments, of the conditional distribution of the hate score.

Table 8: Results from Linear Conditional Quantile Regression

	Coef. (Std. Error)
Quantile 0.75	
Constant	5.33 (0.13)***
Counter-speaking	−0.33 (0.14)**
Censored	−0.33 (0.14)**
Extremely censored	−0.33 (0.18) [†]
Quantile 0.90	
Constant	6.33 (0.09)***
Counter-speaking	0.00 (0.12)
Censored	−0.67 (0.20)
Extremely censored	−0.33 (0.20)
Quantile 0.95	
Constant	7.00 (0.20)***
Counter-speaking	−0.33 (0.36)
Censored	−0.67 (0.23)
Extremely censored	−0.33 (0.23)
Quantile 0.99	
Constant	8.33 (0.42)***
Counter-speaking	0.00 (0.59)
Censored	−1.33 (0.50)**
Extremely censored	−1.00 (0.44)*

Notes:: Linear conditional quantile regressions at the 0.75, 0.90, 0.95 and 0.99 quantiles of the hate score distribution. Quantile regression coefficients are listed with standard errors in parentheses (Obs=1469). Significance levels: *** $p < 0.000$, ** $p < 0.01$, * $p < 0.05$, [†] $p < 0.1$, for a two-sided test

In [Table 9](#) we present three further models of hostile comments (more than 7 in the hate score). Model 1 shows the results for a logistic regression model with clustered standard errors at the individual level (180 clusters). Model 2 shows a logistic regression model with bootstrapped clustered standard errors. The replications in model 2 are based on 180 clusters in the data. Both models show that the effect of the censored treatment is robust, even after taking into account the nested structure of the comments. Model 3 is a rare events logistic regression model. This model, which corrects for small-sample bias, also supports the reduction in the number of hostile comments in the censored treatment.

Table 9: Results of Logistic Regression Results with clustered standard errors (model 1), Logistic regression with bootstrapped clustered standard errors (model 2), and Rare Events regression (model 3) of hostile comments (more than a 7 in the hate score)

	Model (1)	Model (2)	Model (3)
Constant	-2.58 (0.36)***	-2.68 (0.41)***	-2.66 (0.21)***
Counter-speaking	-0.30 (0.50)	-0.30 (0.60)	-0.29 (0.32)
Censored	-1.85 (0.70)**	-1.85 (0.73)*	-1.75 (0.64)**
Extremely censored	-0.30 (0.50)*	-0.41 (0.73)	-0.40 (0.34)
Log Likelihood	-245.13	-245.13	
AIC	498.26	498.26	498.26
Obs.	1469	1469	1469
Pseudo- R^2	0.0341	0.0341	

Notes: Notes: Logistic regression with clustered s.e at the individual level (Model 1), logistic regression with bootstrapped clustered s.e at the individual level (Model 2), and rare events logistic regression (Model 3) with hostile comments as dependent variable. Standard errors in parentheses. Replications of model 2 based on 180 clusters in the data. One or more parameters could not be estimated in 24 bootstrap replicates; standard-error estimates in Model 2 include only complete replications. Significance levels: *** $p < 0.000$, ** $p < 0.01$, * $p < 0.05$, $^\dagger p < 0.1$, for a two-sided test

The estimate for the extremely censored condition is only significantly larger than the estimate for the censoring treatment at $p < 0.10$.

B.2 Robustness Checks

Table 10 shows the results of rare events logistic regression of the comments classified as a norm violation. The results from the multilevel random model in Table 6 are robust. The results from this model have to be interpreted with caution because the rare events logistic model does not account for the nested structure of the data.

Table 10: Rare Events Analysis of Norm Violations (Uncivil Comments)

	Model (1)
Constant	-1.91 (0.16)***
Counter-speaking	0.08 (0.22)
Censored	-0.95 (0.28)**
Extremely censored	-0.50 (0.25)*
AIC	943.07
Obs.	1469

Notes: Rare events logistic regression estimates (Model 1) and logistic regression with clustered standard errors (Model 2) with norm violations as the dependent variable. Standard errors in parentheses. Significance levels: *** $p < 0.000$, ** $p < 0.01$, * $p < 0.05$, $^\dagger p < 0.1$, for a two-sided test

Table 11 shows the results from a multilevel model with two random intercepts (picture and subject) similar to models in table 4. The model shows the interaction effects of the treatments and topic combined after the simple effects for topics and treatments have been taken into account. Poverty is used as the reference category, which means that the interaction terms have to be understood as compared with that category. This model is a reparametrization of model 3 in Table 4.

	Model 1
Main effects	
Constant	4.20 (0.37)**
Counter-speaking	0.07 (0.24)
Censored	-0.02 (0.24)
Extremely censored	-0.12 (0.25)
LGTB	0.22 (0.44)
Refugees/Multiculturalism	0.97 (0.41)*
Feminism	0.19 (0.41)
Interaction effects	
Poverty*Counter-speaking	
LGTB*Counter-speaking	-0.24 (0.25)
Refugees*Counter-speaking	-0.33 (0.24)
Feminism*Counter-speaking	-0.18 (0.23)
Poverty*Censored	
LGTB*Censored	-0.31 (0.25)
Refugees*Censored	-0.62 (0.24)**
Feminism*Censored	-0.30 (0.23)
Poverty*Extremely	
LGTB*Extremely	-0.33 (0.26)
Refugees*Extremely	-0.48 (0.25)
Feminism*Extremely	-0.16 (0.24)
AIC	4371.48
BIC	4472.04
Log Likelihood	-2166.74
Obs	1469
Groups:Subjects	180
Var:Subjects	9
Groups:Subjects	0.44
Var:Subjects	0.11
Residual Variance	0.90

Notes: Linear mixed model fit by REML. Fixed Effects Estimates (Top) and Variance-Covariance Estimates (Bottom) for models of hate speech score. The table lists mean regression coefficient estimates with standard errors in parentheses and p-values calculated based on Satterthwaite’s approximations. Significance levels: *** $p < 0.000$, ** $p < 0.01$, * $p < 0.05$, $^{\dagger}p < 0.1$, for a two-sided test

Table 11: Statistical models

We also validated the model of hate score using a jackknife “leave-one-out” resampling technique. We computed the main model (model 1 in Table 4) leaving out one subject at a time, with a total of 180 models. The distribution of the coefficient estimates of the treatments using this strategy are presented here. Figure 11 shows the estimates in the counter-speaking treatment, Figure 12 shows the estimates for the censored treatment, and Figure 13 shows the estimates for the extremely censored treatment. The value of the coefficients does not change significantly after with the removal

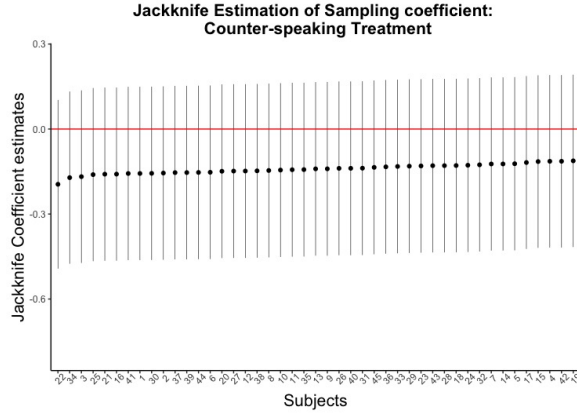


Figure 11: Jackknife Estimation of Sampling Coefficients of the Counter-speaking Treatment

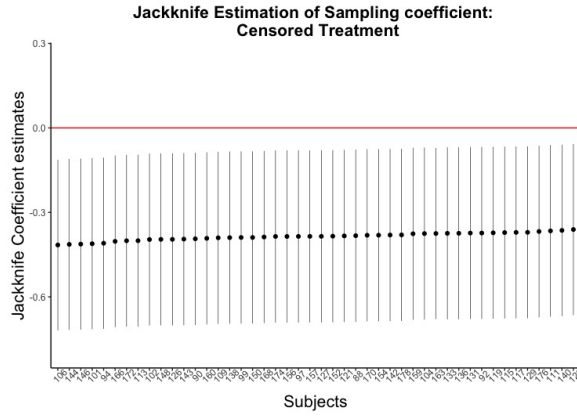


Figure 12: Jackknife Estimation of Sampling Coefficients of the Censored Treatment

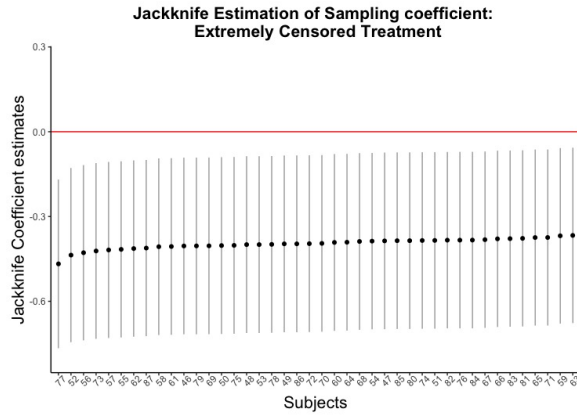


Figure 13: Jackknife Estimation of Sampling Coefficients of the Extremely Censored Treatment

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., & Chavez, A. K. (2013). Norm manipulation, norm evasion: experimental evidence. *Economics and Philosophy*, 29(02), 175-198.
- Bicchieri, C., & McNally, P. (2016). *Shrieking Sirens. Schemata, Scripts, and Social Norms: How Change Occurs* (PPE Working Papers No. 0005). Philosophy, Politics and Economics, University of Pennsylvania.
- Bicchieri, C., & Xiao, E. (2009). Do the right thing: but only if others do so. *Journal of Behavioral Decision Making*, 22(2), 191-208.
- Binns, A. (2012). Don't feed the trolls! managing troublemakers in magazines' online communities. *Journalism Practice*, 6(4), 547-562.
- Blanchard, F. A., Crandall, C. S., Brigham, J. C., & Vaughn, L. A. (1994). Condemning and condoning racism: A social context approach to interracial settings. *Journal of Applied Psychology*, 79(6), 993.
- Blinder, S., Ford, R., & Ivarsflaten, 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(4), 841-857.
- Burgoon, M., Alvaro, E., Grandpre, J., & Voulodakis, M. (2002). Revisiting the theory of psychological reactance. In *The persuasion handbook* (p. 213-232). Thousand Oaks, CA: Sage.
- Bursztyn, L., Egorov, G., & Fiorin, S. (2017, May). *From extreme to mainstream: How social norms unravel* (Working Paper No. 23415). National Bureau of Economic Research. Retrieved from <http://www.nber.org/papers/w23415> doi: 10.3386/w23415
- Chen, D. L., Schonger, M., & Wickens, C. (2016). oTree - an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9, 88-97.
- Cheng, J., Danescu-Niculescu-Mizil, C., Leskovec, J., & Bernstein, M. (2017). Anyone can become a troll. *American Scientist*, 105(3), 152.
- Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annu. Rev. Psychol.*, 55, 591-621.
- Cialdini, R. B., Reno, R. R., & 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(6), 1015.
- Cialdini, R. B., & Trost, M. R. (1998). Social influence: Social norms, conformity and compliance. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol. 2, p. 151-192). New York: McGraw-Hill.
- Citron, D. K., & Norton, H. (2011). Intermediaries and hate speech: Fostering digital citizenship for our information age. *BUL Rev.*, 91, 1435.
- Coleman, J. S. (1990a). The emergence of norms. In M. Hechter, K.-D. Opp, & R. Wippler (Eds.), *Social institutions: Their emergence, maintenance, and effects* (p. 35-39). New York: Aldine de Gruyter.
- Coleman, J. S. (1990b). Norm-generating structures. *The limits of rationality*, 250-273.
- Crandall, Eshleman, & O'Brien. (2002). Social norms and the expression and suppression of prejudice: the struggle for internalization. *Journal of Personality and Social Psychology*, 82(3), 359.
- Crandall, C., & Stangor, C. (2005). Conformity and prejudice. In J. F. Dovidio, P. Glic, & L. Rudman (Eds.), *On the nature of prejudice: Fifty years after allport* (p. 295-309). Malden, MA: Blackwell Publishing.
- DellaVigna, S., List, J. A., Malmendier, U., & Rao, G. (2016). Voting to tell others. *The Review of Economic Studies*, 84(1), 143-181.

- Dovidio, J. F., & Gaertner, S. L. (1986). *Prejudice, discrimination, and racism: Historical trends and contemporary approaches*. Academic Press.
- Duckitt, J. H. (1992). Psychology and prejudice: A historical analysis and integrative framework. *American Psychologist*, 47(10), 1182.
- Elster, J. (1989). Social norms and economic theory. *The Journal of Economic Perspectives*, 3(4), 99-117.
- European Commission against Racism and Intolerance. (2016, March). *Recommendation no. 15 on combating hate speech, adopted on december 2015* (General Policy Recommendation). Strasbourg: Council of Europe.
- Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *The American economic review*, 90(4), 980-994.
- Ford, R. (2008). Is racial prejudice declining in Britain? *The British journal of sociology*, 59(4), 609-636.
- Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). *Countering online hate speech*. UNESCO Publishing.
- Goodman, E., & Cherubini, F. (2013). Online comment moderation: emerging best practices. *Germany: Darmstadt, The World Association of Newspapers WAN-IFRA*. [Http://www.wan-ifra.org/reports/2013/10/04/online-comment-moderation-emerging-best-practices](http://www.wan-ifra.org/reports/2013/10/04/online-comment-moderation-emerging-best-practices) (17.9. 2014).
- Hechter, M., & 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(3), 535-562.
- Henson, B., Reyns, B. W., & 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*, 1043986213507403.
- Hinduja, S., & Patchin, J. W. (2007). Offline consequences of online victimization: School violence and delinquency. *Journal of School Violence*, 6(3), 89-112.
- Ivarsflaten, E., Blinder, S., & 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(4), 421-445.
- Judd, C. M., Westfall, J., & 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., & Steg, L. (2008). The spreading of disorder. *Science*, 322(5908), 1681-1685.
- Kennedy, M. A., & Taylor, M. A. (2010). Online harassment and victimization of college students. *Justice Policy Journal*, 7(1), 1-21.
- Keuschnigg, M., & Wolbring, T. (2015). Disorder, social capital, and norm violation: Three field experiments on the broken windows thesis. *Rationality and Society*, 27(1), 96-126.
- Kraut, R. E., Resnick, P., Kiesler, S., Burke, M., Chen, Y., Kittur, N., ... Riedl, J. (2012). *Building successful online communities: Evidence-based social design*. MIT Press.
- Krupka, E. L., & Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, 11(3), 495-524.
- Mantilla, K. (2013). Gendertrolling: Misogyny adapts to new media. *Feminist Studies*, 39(2), 563-570.
- Matias, J. N. (2016). *Posting rules in online discussions prevents problems & increases participation*.
- Munger, K. (2016). Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, 1-21.
- Paluck, E. L., & 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*, 29-42.
- Pettigrew, T. F. (1991). Normative theory in intergroup relations: Explaining both harmony and conflict. *Psychology & Developing Societies*, 3(1), 3-16.

- Rauhut, H., & Winter, F. (2012). On the validity of laboratory research in the political and social sciences: The example of crime and punishment. In B. Kittel, W. J. Luhan, & R. B. Morton (Eds.), *Experimental political science: Principles and practices* (pp. 209–232). London: Palgrave Macmillan UK. Retrieved from https://doi.org/10.1057/9781137016645_10 doi: 10.1057/9781137016645_10
- Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2016, sep). Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In M. Beißwenger, M. Wojatzki, & T. Zesch (Eds.), *Proceedings of NLP 4 CMC III: 3rd Workshop on natural language processing for computer-mediated communication* (Vol. 17, p. 6-9). Bochum.
- Rost, K., Stahel, L., & Frey, B. S. (2016). Digital social norm enforcement: Online firestorms in social media. *PLoS one*, 11(6), e0155923.
- Schieb, C., & Preuss, M. (2016). Governing hate speech by means of counterspeech on facebook. In *66th ica annual conference, at fukuoka, japan* (pp. 1–23).
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Wadsworth Cengage learning.
- Shapiro, J. R., & 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(4), 877.
- Sherif, M., & Sherif, C. W. (1953). *Groups in harmony and tension; an integration of studies of intergroup relations*. New York: Harper & Brothers.
- Stangor, C., Sechrist, G. B., & Jost, J. T. (2001). Changing racial beliefs by providing consensus information. *Personality and Social Psychology Bulletin*, 27(4), 486-496.
- Tankard, M. E., & Paluck, E. L. (2016). Norm perception as a vehicle for social change. *Social Issues and Policy Review*, 10(1), 181-211.
- Voss, T. (2001). *Game-theoretical perspectives on the emergence of social norms*. na.
- Winter, F., Rauhut, H., & Miller, L. (2017). Dynamic bargaining and normative conflict. *Max-Planck-Institute for Research on Collective Goods Working Paper*..
- Xiao, E., & Bicchieri, C. (2010). When equality trumps reciprocity. *Journal of Economic Psychology*, 31(3), 456-470.
- Zitek, E. M., & Hebl, M. R. (2007). The role of social norm clarity in the influenced expression of prejudice over time. *Journal of Experimental Social Psychology*, 43(6), 867-876.