The Effect of Hate Speech Regulation on Preference Falsification. Experimental Evidence from the US and Germany[*]

Richard Traunmüller (University of Mannheim)[†]

Simon Munzert (Hertie School)

Andrew Guess (Princeton University)

Pablo Barberá (University of Southern California)

JungHwan Yang (University of Illinois at Urbana-Champaign)

This version: November 17, 2020

Abstract. Concerns over the civility of public discourse and the spread of hateful messages have prompted governments into action around the world. Yet, whether and how to restrict speech that is considered offensive or promotes hate toward particular groups remains controversial. Empirically, claims about the expected effectiveness and likely consequences of regulatory intervention are largely untested. We present results from a pre-registered experimental study implemented in the United States and Germany, which tests two unintended consequences of hate speech law on self-censorship of citizens. Our expectations are twofold: First, we hypothesize that restricting the expression of obnoxious ideas merely drives them underground. Second, we expect a chilling effect on public discourse induced by hate speech regulation. The experimental findings run counter to our expectations and provide surprising new insights on the freedom of speech norm and the consequences of restricting hate speech.

Keywords. Hate Speech, Free Speech, Self-Censorship, Preference Falsification, Online, Regulation, USA, Germany, Survey Experiment, List Experiment, Prime

Word count. x,xxx

[*] Previous versions of this paper was presented at EPSA 2019 and 2020. We thank discussants and participants for their questions and comments.

[†] Corresponding author. Address: XXX Mannheim, Germany. Email: xxx, Web: xxx.

Introduction

With the spread of social media platforms and other forms of online communication, many people are exposed to uncivil or hateful communication by members of their network or even complete strangers.[1] Concerns over the civility of public discourse and the spread of hateful messages have prompted governments into action around the world.

Yet, whether and how to restrict speech that is considered offensive or promotes hate toward particular groups remains controversial (Hare and Weinstein, 2009; Bleich, 2011*b*; Herz and Molnar, 2012). Next to struggles over definitions of what constitutes 'hate speech' in the first place,[2] considerable disagreement exists over the adequate regulatory response: how can we limit discriminatory speech while preserving the democratic value of free speech? Answers are difficult because positions for and against the restriction of hate speech are deeply rooted in conflicting normative principles.

Normative issues aside, regulatory efforts to reduce hateful messages can have important effects not only on the spread of hate speech, but on public discourse, discriminatory attitudes, and political behavior. However, claims about the expected effectiveness and likely consequences of regulatory intervention are largely untested (Brown, 2015). As first step to this end, we propose to inform the debate on hate speech regulation by testing some of these claims against experimental evidence. No matter on which side one stands in the normative debate over the benefits and risks of restricting

[1] According to online-representative surveys we conducted in the United States and Germany in 2018, 20% of US respondents reported to have become the target of hate speech online (10% in Germany) and 50% at least witnessed an online hate speech attack (22% in Germany). More details about the survey are reported below.

[2] As law professor Nadine Strossen, a former president of the American Civil Liberties Union, has noted, charges of 'hate speech' are found all over: 'members of the Black Lives Matter movement have been accused of 'hate speech' against police officers, whereas many critiques of the Black Lives Matter movement have been denounced as 'hate speech' against its supporters or against African Americans generally. [...] Evangelical Christians who charge LGBT sexuality as sinful have been accused of 'hate speech' against gay men and lesbians, whereas those who make these charges against evangelical Christians have been accused of religious 'hate speech.' Similarly, critics of some Islamic teachings about women have accused some imams of 'hate speech' against women, whereas these critics have been in turn accused of 'Islamophobia,' or 'hate speech' against Muslims" (see Strossen, 2018, 11-12).

hateful speech, we believe that the discussion is most constructive if it is based on sound evidence of the effects of such legislation.

While the most obvious proposition is that hate speech laws help reduce the spread of hateful communication in the public sphere, our interest here is on the unintended consequences of hate speech legislation. By unintended consequences, we do not mean the mere possibility of abuse of such laws as targeted efforts to restrict free speech for minority groups or those with dissenting political views.[3] We are concerned with the unintended potential harms that directly emanate from hate speech laws themselves and which have been repeatedly put forth by critics of hate speech regulation.

We turn these criticisms of hate speech law into two testable hypotheses. Both hypotheses claim that hate speech regulation leads to self-censoring behavior among citizens.[4] The first hypothesis states that instead of reducing discriminatory ideas, restricting hate speech merely drives them underground (H1). The second hypothesis argues that hate speech regulation results in a more far-reaching 'chilling effect' on public discourse (H2). Thus, the key difference between H1 and H2 is that H1 argues that hate speech law leads to self-censorship in the speech it is designed to stop, whereas H2 states that hate speech law also leads to self-censorship in the speech it is not designed to stop.

Studying self-censorship is tricky because we need to separate what citizens say from what they actually believe. To put these hypotheses to an empirical test, we implemented a design that combines a set of double list experiments with a randomly assigned hate speech regulation prime and a battery of issue preferences that are openly queried. The double list experiments are used to measure true preferences on two potentially sensitive issues in an unobtrusive manner. The battery of openly queried

[3] Note, however, that even in liberal democracies, legal restrictions on free expression (e.g. anti-defamation or copyright protection laws) are sometimes used for alternative political goals such us suppressing media reporting (Stanig, 2015) or silencing dissent (Meserve and Pemstein, 2018).

[4] In the following, we will consistently use the term self-censorship, representing the act of misrepresenting one's wants under perceived social pressures. This is equivalent to Kuran's (1997) definition of preference falsification.

issue preferences is used to elicit openly stated preferences on sensitive issues. Comparing support for those with the preferences elicited by the list experiments allows us to quantify self-censoring behavior. The hate speech regulation prime is used to randomly induce potential effects of hate speech regulation on the disclosure of potentially unpopular opinions via the battery of openly queried issue preferences.

We administered the experiments in online population-representative samples of the US and Germany. A comparison of these two countries allows us to explore the effect of hate speech legislative efforts on self-censorship under variable conditions. The US and Germany not only differ in their cultural tradition concerning the norms of free expression, but also in their actual hate speech legislation (Bleich, 2011*a*; Rosenfeld, 2002; **?**). Whereas no hate speech law exists in the US and its tradition of the First Amendment, Germany legally sanctions certain acts of speech and is a front-runner of the European approach to restricting hate speech (Bleich, 2011*a*).

Our findings provide surprising new insights on the norms of free speech and the consequences of restricting hate speech. We do not find that priming hate speech laws leads to significant self-censoring regarding an offensive statement about Muslims. Interestingly, respondents did not seem to consider the statement particularly sensitive, even though it is based on an actual case of hate speech that lead to a conviction which was upheld by the European Court of Human Rights. This in itself is certainly an interesting finding concerning the relationship between citizens preferences and existing hate speech laws. On a more general level, this null-finding (which is unlikely to be the result of low power or an ineffective treatment) suggests that hate speech legislation may be ineffective in reducing offensive statements about minorities.

Studying citizens' true preference regarding the free expression of unpopular opinions, we also fail to uncover a chilling effect of hate speech regulation. Although freedom of expression is considered a deeply held democratic value, a considerable part of the population actually prefers limits on potentially offensive opinions, which is in line with hate speech regulation. In fact, more people are in favor of restricting free expression than they would be willing to openly admit when directly asked. This suggests the presence of a strong cultural norm of free expression. And counter to previous accounts, the US and Germany do not differ much in this regard. Importantly, hate speech laws have the effect of weakening and qualifying the cultural norm of free expression in the minds of citizens, and thus encourage them to express their actual preference for the restriction of offensive opinions.

The overarching implication of our experimental results is that hate speech laws seem to create their own legitimacy. Once put in place, citizens tend to follow the thrust of these laws and more openly support the restriction of free expression. They are more willing to censor the use of the Internet and more willing to restrict the public expression of unpopular and offensive opinions. Priming hate speech regulation is effective when it reminds people of the actual legislation in place. As a result, this effect is only visible in Germany where hate speech laws exist, but not in the US where such a law is missing.

Hate Speech Regulation in Context

Before turning to the empirical analysis, we consider recent developments in hate speech legislation in liberal democracies and discuss potential intended and unintended consequences for the formation and communication of issue preferences.

Recent Developments in Hate Speech Legislation

States around the world restrict the free expression of their citizens. Driven by their desire to stay in power, authoritarian states censor critical messages as well as calls to collective action that threaten the legitimacy and stability of the regime (Egorov et al., 2009; King et al., 2013, 2014). Democratic politics is not immune from incentives to suppress criticism and dissent (Stanig, 2015; Meserve and Pemstein, 2018). Even liberal democracies that recognize the value of free speech as a constitutional right restrict certain expressions of their citizens. Importantly, many democratic states restrict free speech with the aim of protecting individuals from harmful speech and upholding the public order.

In recent decades, liberal democracies have experienced a rise in legal restrictions on so-called "hate speech" (Hare and Weinstein, 2009). While the term "hate speech" is ill-defined, it is commonly understood as referring to speech that expresses hate towards or discriminatory views about a social group based on the group's identity or characteristics. Hate speech laws differ in the specific groups they protect (e.g., groups defined by race, ethnicity, religion and sexual orientation, or a subset of those) as well as in terms of how they are justified, i.e. by the particular manner or style of speech or by the likelihood of harmful consequences (Post, 2009). In some countries, speech is prohibited that offends, insults or degrades a social group. For instance, Section 18C of the *Australian Racial Discrimination Act* prohibits any act that is "reasonably likely [...] to offend, insult, humiliate or intimidate another person or a group of people" and is "done because of the race, colour or national or ethnic origin of the other person or of some or all of the people in the group." In other instances, speech is prohibited because it is likely to have harmful effects such as violence or discrimination. According to Article 130(1) of the *German Criminal Code*, a person commits a criminal offense if she "in a manner capable of disturbing the public peace [...] incites hatred against segments of the population or calls for violent or arbitrary measures against them" or "assaults the human dignity of others by insulting, maliciously maligning, or defaming segments of the population." As Bleich (2011*a*, 18) concludes, '[t]he variety of legal formulations and protected groups is dizzying, but the overarching logic is similar on one fundamental level – these laws aim to restrict forms of speech that target people because of [their] core identities."

With the rise of the internet as an ecosystem for mass communication, a we observe new generation of hate speech laws that are targeted to combat hateful and offensive content online. The German *Network Enforcement Act* (the 'Netzwerkdurchsetzungsgesetz' or 'NetzDG'), which came into effect in 2018, is a front runner in that regard. This new law requires social media platforms to delete "obviously illegal" content within 24 hours and to review potentially illegal content with a week or face a hefty fine of up to 50 million euros. Similar laws that restrict harmful online content (which may cover more than merely 'hate speech' in the narrow sense, e.g. 'cyber-bullying' or 'fake news') have been passed in Australia (the 2019 'Abhorrent Violent Material Act') and Russia. Similar social media legislation is currently discussed in France, the UK, and the EU.

The Free Speech–Hate Speech Trade-Off

Despite their widespread existence and growth in liberal democracies other than the USA, hate speech laws remain controversial. Proponents of hate speech restrictions point to the harmful psychological or even physical consequences of discriminatory speech (Matsuda, 1989) and stress that the value of free speech is not absolute, but has to be weighed against competing values, such as equality and personal dignity (Fish, 1994; Parekh, 2012; Waldron, 2012). Critics of this view insist that speech is not only quite different from action, but also deserves special protection because of its fundamental importance for human autonomy, democratic self-governance, and political legitimacy (Dworkin, 1999; Post, 2009). Thus, in a sense, the debate of hate speech regulation epitomizes the democratic dilemma of balancing liberty and equality.

Besides disagreement over the effects of hateful speech and the precedence of normative principles, the debate over hate speech regulation is also a debate about the expected effects and the likely consequences of hate speech law. These arguments are of particular interest because they involve claims that are, at least in principle, amenable to empirical testing. It is a curious feature of the hate speech debate that it has been, in large parts, reluctant to provide such empirical evidence.

The criticism of legal restrictions on hate speech and their likely consequences is generally two-fold. A first set of concerns revolves about the idea that, counter to its intended purpose, hate speech law is simply an ineffective measure and unlikely to deter hateful and discriminatory expression in any meaningful way. A version of this argument states that the merely symbolic nature of hate speech laws make them a cheap and more easily adopted alternative to more costly efforts of actually and directly combating the roots of discrimination (Baker 2012, Strossen 2018). Many believe that, instead of legal sanctions, 'counter speech' is a more effective way to respond to hate speech.

A second set of concerns points to the potential dangers of restricting hate speech by legal prohibition to forms of expression that are in fact not hate speech. The most prominent argument is the "slippery slope", i.e. that "once the door to regulation is open ever so slightly it is bound gradually to open wider, eventually allowing for censorship of all kinds of legitimate yet unpopular speech" (Rosenfeld 2012: 286). A related concern is that the inherent vagueness of hate speech laws leaves too much room for interpretation and is therefore likely to lead to false decisions and the unacceptable censoring of minority views or political view points. This is so because "the expression targeted by bans on speech that insult or demean individuals on the basis of race, ethnicity, religion, or sexual orientation is rarely a mere rant against members of these groups but is almost always bound up with criticism of some government policy, e.g., immigration laws or race-based minority preferences.” (Hare and Weinstein, 2009, 5).

In the following, we will link these considerations to empirically observable citizen behavior and derive testable hypotheses.

Attitudinal and Behavioral Consequences of Hate Speech Legislation

Regulation of hate speech impacts the behavior of citizens by changing the costs and benefits associated with the choice between free expression and self-censorship (Kuran, 1997). To avoid the risk of being sanctioned for hateful speech, citizens will avoid expressing their controversial opinion openly and instead chose to self-censor.

Even if convictions are rare and the actual risk of sanctions is low, hate speech laws may also lead to self-censoring behavior because they signal the boundaries of socially acceptable speech. According to Parekh (2012, 46), hate speech legislation “lays down norms of civility and sends out clear messages concerning what is or is not an acceptable way of talking about and treating other members of society. Being a collective and public statement of the community’s moral identity and guiding values, the law affirms and enforces these values, has a symbolic and educational significance, and helps shape the collective ethos.” Thus, citizens may also falsify their true preference because of their concern with social consequences and their desire to conform to socio-cultural norms (Loury, 1994). Scholars have repeatedly stressed that social pressures are at least as likely to prevent citizens from freely speaking their mind as legal sanctions (Mill, 2011; Gibson, 2006).

The first hypothesis holds that instead of pushing back obnoxious ideas, restricting hate speech merely drives them underground and thus results in an increase in self-censorship (H1). Faced with sanctions for hate speech, people will show a greater difference between what they truly believe and what they say they believe with regards to a social group protected under hate speech legislation. This behavioral effect would be highly problematic because it would generate additional societal costs: when people who hold problematic views do not express them openly, it is more difficult to know who holds these views and how popular they are. Moreover, their ideas cannot be publicly challenged (Strossen, 2018).

The second hypothesis states that, regardless of its effect on true hate speech, hate speech regulation results in a chilling effect on public discourse (H2) more generally. Here, “the concern is that in practice any attempt to ban hate speech, no matter how careful or well-intentioned, is bound to cause collateral damage in terms of making people overly cautious about what they will say in public” (Brown, 2015, 266). Faced with the uncertainty or vagueness of hate speech regulation, citizens are reluctant to openly debate controversial, yet crucial political issues and thus less willing to reveal their policy preferences. This effect is particularly harmful for democracy because democratic governance rests fundamentally on the free debate of policy options.

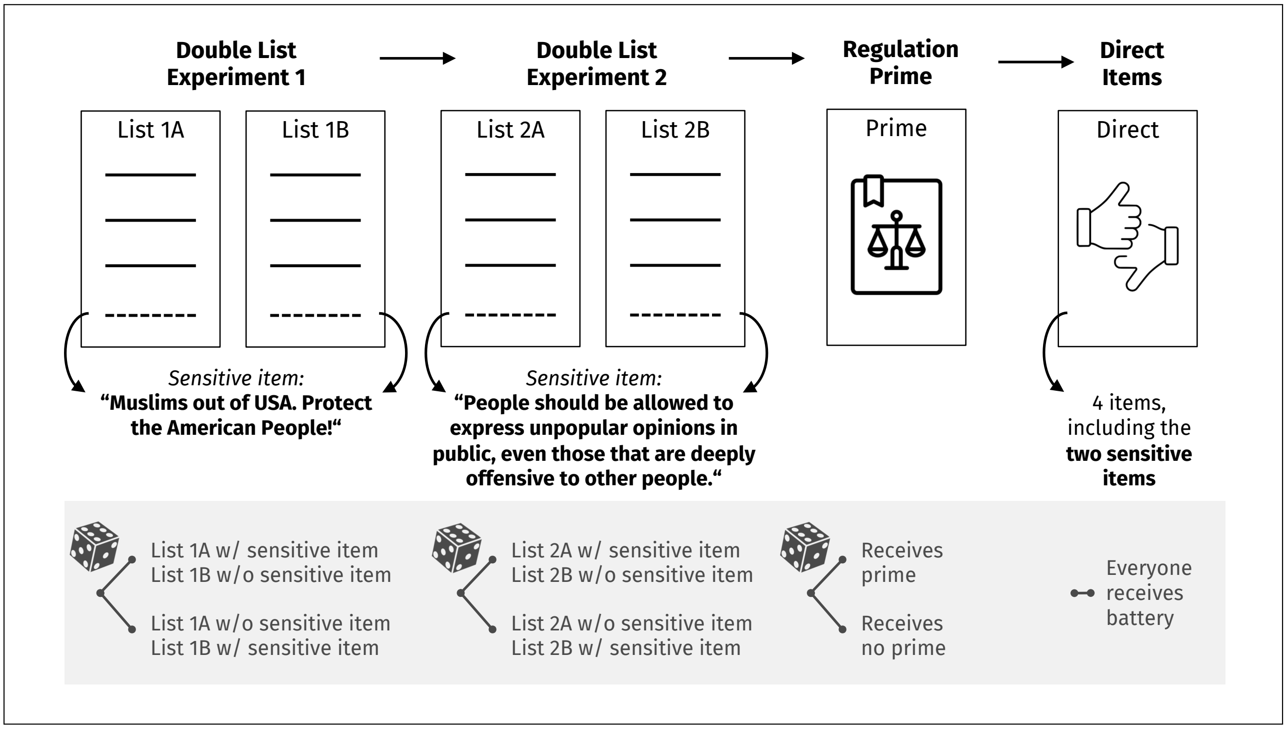
Third, we expect **country differences** in the baseline levels of preferences as revealed in the list experiments. To put our hypotheses to an empirical test, we embed our experimental design in surveys in the United States and Germany, two settings for which the cultural tradition concerning free expression differs a lot (First Amendment vs. continental tradition). A global survey has recently found strong variation in the extent citizens in 64 countries support free expression (Wike and Simmons, 2015). On an index from 0 to 8 (least to most supportive of free expression), US citizens ranked highest with a mean of 5.73, Germans considerably lower with only 4.34.

However, the countries also differ in their legal approach to regulating hate speech. Whereas no hate speech law exists in the US, Articles 86 and 86a of the German Criminal Code, prohibit the spread of Nazi propaganda and symbols and Article 130 prohibits incitement of the people ('Volksverhetzung') and Holocaust denial. As recently as in 2018, Germany has introduced a new and controversial social media law to combat online hate speech (the so-called "Netzwerkdurchsetzungsgesetz"), the first of its kind on the globe. Therefore, we expect higher baseline revealed support (in the list experiment) for the free speech statement in the US compared to Germany. We have no prior expectations regarding baseline differences in support for the offensive statement toward Muslim immigrants. However, we have no a priori expectations about country differences in the effect of hate speech regulation. We expect H1 and H2 to hold in both contexts.

Experimental Setup

In order to put the hypothesized effects of hate speech regulation on self-censorship to an empirical test, three elements are key: (1) true preferences on sensitive issues, (2) openly stated preferences on sensitive issues, and (3) hate speech regulation that potentially affects differences between (1) and (2). To capture these elements, we implemented a design that combines a set of double list experiments with a randomly assigned hate speech regulation prime and a battery of issue preferences that are openly queried. The double list experiments are used to measure true preferences on two potentially sensitive issues in an unobtrusive manner. The battery of openly queried issue preferences is used to elicit openly stated preferences on sensitive issues. The hate speech regulation prime is used to randomly induce potential effects of hate speech regulation on the disclosure of potentially unpopular

Figure 1: Illustration of experimental setup


Double List Experiment 1
Double List Experiment 2
Regulation Prime
Direct Items
List 1A
List 1B
List 2A
List 2B
Prime
Direct
Sensitive item:
"Muslims out of USA. Protect the American People!"
Sensitive item:
"People should be allowed to express unpopular opinions in public, even those that are deeply offensive to other people."
4 items, including the two sensitive items
List 1A w/ sensitive item
List 1B w/o sensitive item
List 1A w/o sensitive item
List 1B w/ sensitive item
List 2A w/ sensitive item
List 2B w/o sensitive item
List 2A w/o sensitive item
List 2B w/ sensitive item
Receives prime
Receives no prime
Everyone receives battery

opinions via the battery of openly queried issue preferences. Figure 1 provides an schematic illustration of the experimental setup, which we will explain in more detail in the following.

Priming Hate Speech Law

After multiple buffer items, we primed half of the respondents with a fictitious hate speech law.[5] The intention was to activate norms of civility and the boundaries of acceptable expression. Thus, our experiment is designed to capture the educative or symbolic effect of hate speech regulation.[6]

We used fictitious legislation to ensure the comparability between the US and the German sample in the sense that respondents from both countries are equally confronted with a law unknown to

[5] The list experiments were placed at the beginning of the survey, the prime together with the direct items were placed at the end of the survey, with six (German survey) and twelve (US survey) buffer items in between to avoid potential interference (Eady, 2017).

[6] We regard it as unlikely that the hate speech prime actually instilled any concern for legal consequences of in-survey behavior in the respondents' minds.

Figure 2: Design of Hate Speech Regulation Prime

> As you may have heard, the government is making serious efforts to combat online hate speech. This could mean that a large number of social media posts with offensive or hateful content will be deleted and legally prosecuted. The content of hate speech legislation that is currently discussed is described in the following text. Please read it very carefully and make sure you understand it.
>
> "A person is guilty of an offense if she sends a message over an online platform which
>
> - uses threatening, abusive or insulting words, or
> - displays any writing, image or video which is threatening, abusive or insulting,
>
> if she intends thereby to stir up hatred against a religious group.
>
> A person guilty of an offense under this law is liable for a prison term not exceeding six months or a fine or both.
>
> This law does not prohibit or restrict discussion, criticism or expressions of antipathy, dislike, ridicule, insult or abuse of particular religions or the beliefs or practices of their adherents."

them. However, it closely resembles actual legislation in the UK, combining fragments from the Public Order Act 1986, the Communications Act 2003 and the Racial and Religious Hatred Act 2006 of the Parliament of the United Kingdom. Yet, we simplified the legalistic language to reduce the cognitive burden on the respondents.

To minimize deception, we did not suggest that such a law was in place but told respondents that the law is currently discussed.[7] After instructing them to carefully read the text, we ask whether respondents oppose or support the law on a five-point scale. The wording of the prime is documented in Figure 2 (see Figure D7 in Online Appendix D for the wording in the German survey). The assignment of the primed and non-primed group is blocked on the variables gender, age, and education. Balance checks suggest the randomization worked.

To check whether respondents actually take the time to carefully read the fictitious hate speech law, we ran an attention check just before the experimental manipulation (Berinsky et al., 2012). In this attention check, respondents were asked to ignore the initial question (about the switch from daylight

[7] Although Germany already has a new social media law targeting hate speech (NetzDG) that came into effect in early 2018, it is still discussed in the public and among policymakers.

saving time) and to just type 'read' into the open text field. In the US sample, 89 percent passed the attention check, while in the German sample, 85 percent passed. As a further check, we timed the duration the respondents spent reading the fictitious hate speech law. In the US sample, the average reading time was XX seconds (s.d. = XXX, min = XXX, max = XXX) and in the German sample 64 seconds (s.d. = 100, min = 1.6, max = 1563).

Among those who passed the attention check in the German sample, a majority rather support (33 percent) or strongly support (20 percent) the hypothetical hate speech law. Only a small minority rather oppose (9 percent) or strongly oppose it (7 percent). In the US sample, support is somewhat lower with 9 percent strongly and 28 percent rather supporting it. On the other hand, 21 percent report to strongly oppose and 19 percent report to rather oppose it. On the surface, these results are in line with differences in societal norms and existing legislation on free speech and hate speech legislation between both countries.

Measuring True Preferences: Two Double List Experiments

Two double list experiments are used to measure people's willingness or reluctance to openly express controversial or discriminatory views. While this unobtrusive method is intended to avoid the ubiquitous problem of social desirability bias in sensitive survey questions (Tourangeau and Yan, 2007), we rely on it to elicit survey item misreporting as a quantity of substantive interest (Aronow et al., 2015; Eady, 2017; Gilens et al., 1998). It is precisely the misreporting of sensitive opinions that lets us learn about the presence of self-censoring behavior.

In a double list experiment, both the treatment and the control group get two lists of items. For the first group, the sensitive item is included in the first list, for the second group, the sensitive item is included in the second list. After considering each list, respondents are asked to report how many of the statements (or posts) they support. But just reporting the number, it is not possible to identify individuals' support for particular items.[8] We then obtain two estimates of the sensitive item: one estimate from comparing the outcomes of the first list and the second from comparing the outcomes of second list. Averaging across these two estimates produces the final estimate. Importantly, the variance of this combined estimator is considerably smaller than for the single estimators (Glynn, 2013).

The first sensitive item we use is "Muslims out of USA (Germany). Protect the American (German) People!"[9] This item was modeled after a prominent hate speech case in the UK (Norwood v Director of Public Prosecutions, 2003). Mark Norwood, a regional organizer of the British National Party (BNP), was convicted and fined for displaying a poster with the slogan "Islam out of Britain. Protect the British people." It was found to be abusive and insulting to Muslims, to cause "harassment, alarm, and distress" and to be motivated by religious hate (see Hare and Weinstein, 2009, for a critical discussion of this case). We decided to change the wording from "Islam" to "Muslims" to make the statement a clearer case of hate speech by targeting a group of individuals based on their identity and not simply a religious faith. This item will be used to test H1, i.e. the expectation that hate speech legislation leads to self-censorship on sensitive items. Specifically, we expect the difference between the indirectly and directly revealed agreement rates towards this item to be larger in the prime group. By mimicking an authentic hate speech case, we are able to put our findings in perspective of actual hate speech legislation, which is, however, unlikely to have had a direct impact on public discourse in the contexts under study. At the same time, resentment against Muslims is prevalent in both countries, making it a relevant statement.[10]

[8] This is not true, of course, when respondents choose 0 or report to support all items in the list.

[9] The original German wordings are reported in Online Appendix D.

[10] In the Spring 2016 edition of the Pew Global Attitudes survey, 29% of German respondents openly reported to have a mostly unfavorable or very unfavorable opinion of Muslims in Germany (Pew Research Center, 2016). Similarly, in Pew

The second sensitive we use is “People should be allowed to express unpopular opinions in public, even those that are deeply offensive to other people.” This item operationalizes agreement with the free speech norm under adversarial circumstances. Thereby, it abstracts away from concrete hate speech and focuses on what should or should not be allowed in public discourse on a more general note. This item will be used to test H2, i.e. the expectation that hate speech legislation induces a chilling effect on public discourse. Specifically, we expect hate speech legislation to induce lower public agreement with the free speech norm, leading to larger differences difference between the indirectly and directly revealed agreement rates towards this item in the prime group.

To ensure that the variance of the estimator is minimized, we have to carefully design the control items. A particular challenge is to balance the use of low (or high) prevalence items that reduce the variance but invite the possibility of floor (or ceiling) effects. One solution is to use negatively correlated items within a list and positively correlated lists (Glynn, 2013). Our strategy is to design two lists of $J = 3$ items each. To select these items, we rely on items that were evaluated in a pretest using a sample of U.S. MTurk workers. Table 1 reports the full wordings of all four lists.

The bottom part of Figure 1 illustrates the assignment procedure. For the first list pair 1A and 1B, respondents are randomly assigned to receiving the sensitive item S_1, “Muslims out of USA (Germany). Protect the American (German) People!” in either list 1A or list 1B. For the second list pair, respondents are again randomly assigned to receiving the sensitive item S_2, “People should be allowed to express unpopular opinions in public, even those that are deeply offensive to other people.” in either list 2A or list 2B. This assignment is independent of the assignment of the sensitive item in the list pair 1A/1B.

Research Center’s American Trends Panel Poll fielded in February 2019, 25% of respondents rated Muslims 33 or lower on a 100-degree feeling thermometer (Pew Research Center, 2019).

In addition, we randomize the order of the items within each list. The assignment of the treatment and control lists is blocked on the variables gender, age, and education. Tables A2 and A3 in the Online Appendix provide balance tests for double list experiment 1 and 2, respectively.

Measuring Public Preferences: Direct Items

In order to contrast the support for the two sensitive items elicited indirectly using double list experiments with the answers given in direct questioning, we include a short battery of direct items. Two of these direct items exactly match the sensitive items used in the list experiments: "Muslims out of USA (Germany). Protect the American (German) People!" and "People should be allowed to express unpopular opinions in public, even those that are deeply offensive to other people." These will allow us to quantify the degree of self-censorship by comparing the answers of the list experiment to the answers to direct survey items. In addition, we add another item on religious freedom ("People should be able to practice their religion freely in our country.") and another item on free speech on the internet ("It is important that people can use the Internet without government censorship."). The direct item battery immediately followed the prime or the attention check for those who are members of the control group (see again Figure 1). For each of the items, respondents were asked whether they supported or opposed the statement.

Data and Sample

We embedded our experiment in two panel surveys fielded on initially about 1,500 respondents recruited for the YouGov U.S. Pulse panel and the YouGov German Pulse panel each, which enables

Table 1: Instructions and items in the double list experiments

List	Instruction	Items
1A	In the following you will find statements posted on social media that some people support or oppose. After you have read all statements, please tell us HOW MANY of them you support. We do not want to know which ones, just HOW MANY.	• “Marijuana is harmless. Legalize it!” • “Vaccines save lives. Mandatory vaccination for all children!” • “Stop killing babies. Make Abortion illegal in ALL cases!” • “Muslims out of USA. Protect the American People!” **[Treatment item]**
1B	Here is another set of statements posted on social media that some people support or oppose. Please tell us again HOW MANY of them you support. We do not want to know which ones, just HOW MANY.	• “More Women in Tech. Affirmative Action now!” • “Guns don’t kill people. People kill people!” • “‘Save the planet. Raise the taxes on gasoline!” • “Muslims out of USA. Protect the American People!” **[Treatment item]**
2A	In the following you will find statements that some people support or oppose. After you have read all statements, please tell us HOW MANY of them you support. We do not want to know which ones, just HOW MANY.	• “People should be able to make statements that criticize the government publicly.” • “Media organizations should be able to publish information about large political protests in our country.” • “Government should be able to stop a news media outlet from publishing biased or inaccurate information.” • “People should be allowed to express unpopular opinions in public, even those that are deeply offensive to other people.” **[Treatment item]**
2B	In the following you will find statements posted on social media that some people support or oppose. After you have read all statements, please tell us HOW MANY of them you support. We do not want to know which ones, just HOW MANY.	• “Marijuana should be legalized, even if it may be harmful for some people.” • “Vaccination should be mandatory for all children, even if parents oppose it.” • “Abortion should be illegal, even if there is a health risk for the mother.” • “People should be allowed to express unpopular opinions in public, even those that are deeply offensive to other people.” **[Treatment item]**

tracking of people's web usage on desktop and mobile devices. The Pulse panel is a subset of YouGov's traditional survey panels, where respondents opt in to install tracking software on their devices.

In both surveys, panelists that installed the web tracking software RealityMine on their computers and cell phones agreed to participate in a "Politics and Media" study with multiple survey waves. Their participation was rewarded using YouGov's proprietary point system and included a bonus if the respondent completed all waves in order to disincentivize attrition. Participation was voluntary and respondents were able to opt-out from the web tracking part of the study at any point in time. Respondents were sampled using age, gender, party identification, and education quotas and then re-weighted in order to obtain a sample that is representative of the U.S. population on these characteristics.

The wave in which the list experiment was embedded was fielded to $N_{GER} = 1,429$ respondents between Dec 6, 2018 and Dec 21, 2018 in Germany. The setup in the US survey was initially implemented incorrectly twice. In the first survey (wave 5 of the original panel fielded between Dec 20, 2018 and Jan 7, 2019) none of the respondents in the list experiments received the sensitive items, while prime randomization was carried out correctly. In the second survey (wave 6 of the original panel, which replicated wave 5, and which was fielded between Jan 24 and Feb 5, 2019), the lists were implemented correctly, but all respondents received the prime. In another survey, which went to a new sample of $1,506$ respondents and was fielded between Jun 6 and Jun 29, 2020 (but otherwise replicated the original questionnaire as far as possible), both components were implemented correctly. For the analyses reported in the main text, we pool the new sample with the respondents participating in both Waves 5 and 6, giving us a total sample size of $N_{USA} = 2,806$ respondents. To make use of the Wave 5/6 data, we combine the properly implemented law prime treatment status and outcomes on the direct items from Wave 5 with the results from the properly implemented list experiment from Wave 6.

Although this was not planned as such, these repeated runs give us the opportunity to increase the overall power of the experiment on the one hand and to perform several additional tests to verify the robustness of the results on the other. Section A in the Online Appendix provides a more detailed overview of the survey and experimental setup.

Quantities of Interest

Based on our research design, we are able to identify several quantities of interest (see Table 2). The key design idea of our project is a systematic comparison of indirect and direct question items, i.e. Y_{true} and Y_{reported}. This allows us to elicit a behavioral expression of *self-censorship*, which will serve as the main outcome in our study:

$$\text{Self-censorship} = \text{Y}_{\text{true}} - \text{Y}_{\text{reported}}$$

We refer to preference revealed in the double list experiment as the "true" preference. Two assumptions are necessary to justify this (Blair and Imai, 2012). First, we have to assume that subjects respond truthfully to the list experiment ("no liars"), which unfortunately cannot be tested. Second, we have to assume that the inclusion of the sensitive item does not affect the answer to the control items ("no design effect"). This assumption can be tested and all list experiments in our study fail to reject the null of no design effect (see Tables C10 and C11 in the Online Appendix). The answers to the direct items are considered the reported preferences. It is important to remember that we are only able to identify self-censoring behavior on the group, not the individual level.

Our research design allows us to not only assess whether preference falsification occurs and to what degree, but to manipulate this self-censoring behavior in a priming experiment. It is this option that we will leverage to study the effects of hate speech regulation. Comparing respondents who re-

Table 2: Identification of quantities of interest

		Experimental condition			
		Prime	**No Prime**	**Prime − No Prime**	**Quantity of interest**
Preference measure	**Indirect**	$Y^{\text{prime}}_{\text{true}}$	$Y^{\text{no.prime}}_{\text{true}}$	$Y^{\text{prime}}_{\text{true}} - Y^{\text{no.prime}}_{\text{true}}$	*(a) Difference in true preference*
	Direct	$Y^{\text{prime}}_{\text{reported}}$	$Y^{\text{no.prime}}_{\text{reported}}$	$Y^{\text{prime}}_{\text{reported}} - Y^{\text{no.prime}}_{\text{reported}}$	*(b) Difference in reported preference*
	Indirect − Direct	$Y^{\text{prime}}_{\text{true}} - Y^{\text{prime}}_{\text{reported}}$	$Y^{\text{no.prime}}_{\text{true}} - Y^{\text{no.prime}}_{\text{reported}}$	$(Y^{\text{prime}}_{\text{true}} - Y^{\text{prime}}_{\text{reported}}) -$ $(Y^{\text{no.prime}}_{\text{true}} - Y^{\text{no.prime}}_{\text{reported}})$	*(c) Difference in self-censorship*

ceived the prime to those who did not, we are able formulate three causal quantities of interest: (a) the *difference in true preferences*, (b) the *difference in reported preferences*, and (c) the *difference in preference falsification*.

As the priming experiment comes after the double list experiment in our setup, the true preferences are pre-treatment variables. We therefore cannot estimate a causal effect of the prime on true preferences. Instead, we expect to find no difference (assuming that the randomization worked and the primed and non-primed groups are balanced). However, we are able to identify a causal effect of priming hate speech legislation on reported preferences. Our main quantity of interest is how the primed and non-primed groups differ in terms of preference falsification. This quantity is a difference-in-differences (see bottom-right corner in Table 2) and is the core outcome of our main hypotheses. For H1, we will rely on the sensitive "Muslim" item, for H2, we will use the sensitive "Freedom of expression" item, respectively. In both instances, we expect the effect of the prime on the difference in preference falsification to be positive, i.e. the prime leading to lower rates of reported support of sensitive items.

Results

Establishing Base Rates of Preference-Falsification

In a first step, we will establish the extent of preference-falsification related to the two sensitive items: the offensive Muslim statement and the free speech preference. To do this, we first calculate the prevalence of the sensitive items using a simple differences-in-means estimator between treatment and control groups, averaging over the two lists of each item. The level of preference-falsification is derived by the difference between this prevalence and the support for the matching direct items. We calculate these key quantities separately for Germany and the US and test for differences between the two contexts.

Using the double list experiment, we estimate that 18 percent of the respondents in the US sample support the statement "Muslims out of USA. Protect the American people" (see Table 3). Support is significantly higher in the German sample, where an estimated 27 percent are in favor of the statement "Muslims out of Germany. Protect the German people." However, this statement does not appear to be particularly sensitive in either the US or Germany. 20 percent of Americans openly support the respective direct item and so do 30 percent of the Germans. The difference between the indirect and the direct item is small and not significant, suggesting that there is not much self-censoring with regard to this statement. This is remarkable given the fact that the sensitive item mirrors an actual case that lead to a conviction under British hate speech legislation and that was upheld by the European Court of Human Rights (Norwood v Director of Public Prosecutions, 2003). However, in a recent review, Blair et al. (2018) found nearly no evidence of sensitivity bias in measures of prejudice.

The second double list experiment yields an average support for the statement "People should be allowed to express unpopular opinions in public, even those that are deeply offensive to other peo-

Table 3: Support rates of statement "Muslims out of USA/Germany. Protect the American/German people!" (double list experiment 1)

		USA		Germany	
		Est.	SE	Est.	SE
List 1a	Mean number of items, treatment	1.71	0.02	1.65	0.03
	Mean number of items, control	1.57	0.02	1.39	0.03
	Prevalence of sensitive item	0.14	0.03	0.26	0.04
List 1b	Mean number of items, treatment	1.55	0.02	1.69	0.04
	Mean number of items, control	1.39	0.02	1.41	0.03
	Prevalence of sensitive item	0.16	0.03	0.28	0.05
Combined	Prevalence of sensitive item, list	0.15	0.02	0.27	0.03
	Prevalence of sensitive item, direct	0.21	0.01	0.30	0.01
	Difference (self-censorship)	-0.05	0.02	-0.03	0.03

ple" of 67 percent among US respondents (see Table 4). It is interesting to contrast this finding with the sample from Germany, where speech norms are generally thought to be different and freedom of speech less sacrosanct. However, we find very similar levels of support among German respondents (66 percent). Interestingly, this indirectly elicited support for free speech is *lower* than the support expressed in the direct questions, where 80 percent of the American and 78 of the Germans claim to support this statement. The difference is sizable and statistically significant in both samples (United States: -13 percent, Germany: -11 percent). This finding suggests that respondents feel compelled to take a stronger pro free-speech stance than they actually believe in. While a broadly shared cultural norm of free speech is what one would expect in the US context, it is interesting that the German context is actually not at all that different.

The Effect of Hate Speech Legislation on Reported Preferences

Next, we assess whether being primed with the hate speech law has an effect on respondents' expressed support for the four items on religion and freedom of speech (corresponding to quantity of interest (b)

Table 4: Support rates of statement "People should be allowed to express unpopular opinions in public, even those that are deeply offensive to other people." (double list experiment 2)

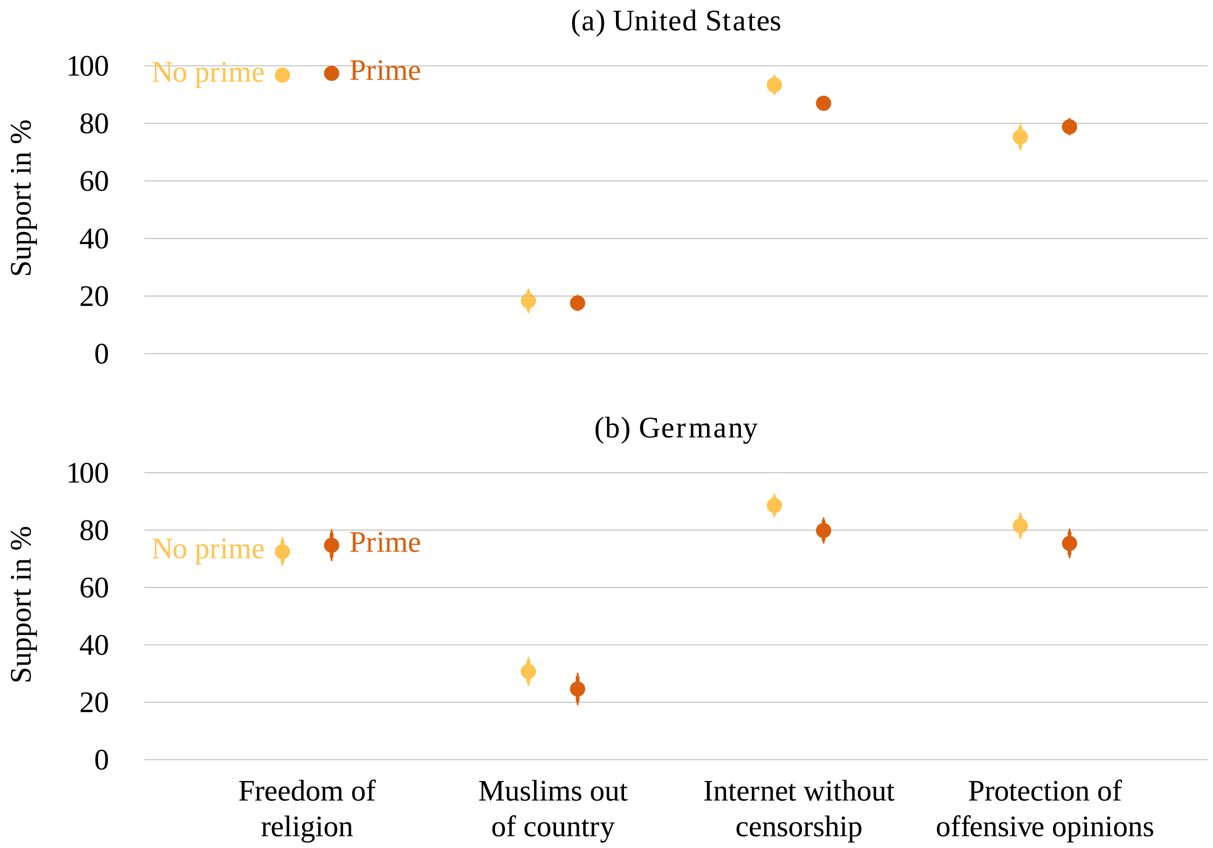
		USA		Germany	
		Est.	SE	Est.	SE
List 2a	Mean number of items, treatment	2.63	0.03	2.84	0.04
	Mean number of items, control	1.99	0.02	2.12	0.03
	Prevalence of sensitive item	0.64	0.03	0.72	0.05
List 2b	Mean number of items, treatment	2.06	0.02	1.88	0.03
	Mean number of items, control	1.40	0.02	1.27	0.03
	Prevalence of sensitive item	0.66	0.03	0.61	0.04
Combined	Prevalence of sensitive item, list	0.65	0.02	0.66	0.03
	Prevalence of sensitive item, direct	0.77	0.01	0.78	0.01
	Difference (self-censorship)	-0.11	0.02	-0.11	0.03

in Table 2). Two of these items are the sensitive items also used in the list experiments. In addition, we added an item on religious freedom and an item on internet censorship. We report the effects (group mean differences) for all respondents who passed the attention check (89 percent in the US sample and 85 percent in the German sample) in Figure 3 (see Tables C12 and C13 in the Online Appendix for numeric results).

We find that priming hate speech regulation does not affect the expression of statements on religious freedom. The notion that people should be able to practice their religion freely is an undisputed believe in the US (98 percent in our sample support the statement) and not moved by priming hate speech law. The results in the German sample suggest that this is not merely due to a ceiling effect. In Germany, only three quarters support religious freedom, and even with this markedly lower support, priming hate speech law has no effect.

The effect of priming hate speech law on the open support of the controversial item "Muslims out of USA/Germany. Protect the American/German people!" does not reach the pre-determined 5% level of statistical significance in either the US or German sample. This is in itself an important result

Figure 3: Effect of priming hate speech legislation on direct support rates of various statements, United States and Germany



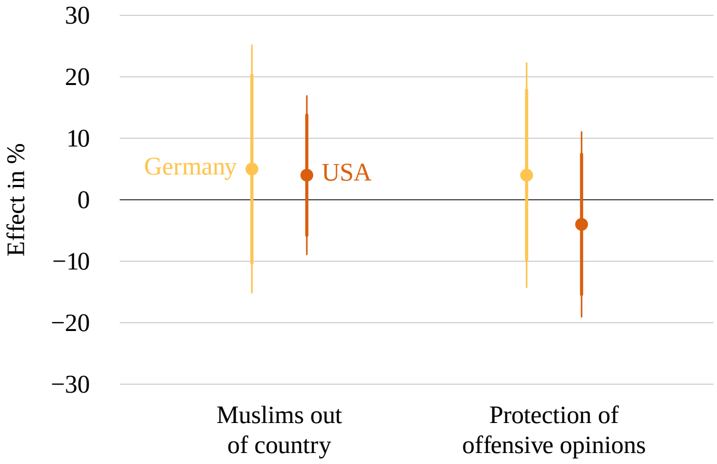
because it suggests that informing people of hate speech legislation and prevailing norms of civility does not necessarily induce a change in their stated opinion concerning the protected group. However, we still want to report on some interesting trends that emerge in the data. Based on the more lenient 10% error probability, respondents the German sample are 4 percentage points less likely to openly support the controversial statement on Muslims, pointing in the hypothesized direction. In the US sample, however, priming hate speech law actually slightly increases the support of the controversial item by 4 percentage points ($p < .1$) in what could be called a 'backlash effect'. Indeed, this is another possible unintended consequence of hate speech regulation. Overall, however, we fail to detect consistent and strong effects of priming hate speech legislation on open support on the religious items.

In contrast, we do find that the prime affects support for the statements more directly related to free expression. Interestingly, German respondents are less supportive of free expression when primed with a fictitious hate speech law. They are significantly less likely to say that "it is important that people can use the Internet without censorship" (-7 percentage points, $p < .01$) and significantly less likely to state that "people should be allowed to express unpopular opinions in public, even those that are deeply offensive to other people" (-5 percentage points, $p < .05$). This suggests that hate speech legislation has the potential to change public opinion about the limits of free expression, because people tend to follow the thrust of the law. This is in accordance with what we know about how citizens reason about civil liberties (Chong, 1993). Legal norms are an important frame of reference. Awareness of these norms generally promotes support in their favor. However, we do not find this pattern in the US sample, which might suggest that priming hate speech law is ineffective if no such law is actually in place.

The Effect of Hate Speech Legislation on Preference Falsification

Our main quantity of interest that captures the hypothesized unintended consequences of hate speech law is the difference in self-censoring behavior between those primed with hate speech legislation vs. those that were not primed (corresponding to quantity of interest (c) in Table 2). This is essentially a difference-in-differences estimate. We again report the results for all respondents who passed the attention check in Figure 4 (see Tables C15 and C14 in the Online Appendix for numeric results).

Figure 4: Effect of priming hate speech legislation on self-censorship, United States and Germany



We can quickly summarize the main finding: we fail to uncover any significant self-censoring effects of priming hate speech legislation. In both the US and the German sample, respondents are slightly more likely to self-censor their statement about Muslims if they do *not* receive a prime than when they do. Regarding the preference for the free expression of offensive speech, US respondents are more likely to self-censor when primed with a hate speech law. German respondents, on the other hand, are less likely to self-censor and instead more likely to reveal their true preference. But these effects are small and not statistically reliable.

It is well known that list experiments are inefficient (i.e. they produce high variances) and even though the double list experiment increases efficiency, we may still not have enough power to reject the null of no effect. Indeed, a power analysis which we conducted on the basis of pre-test data suggests that we may need larger samples to reliably detect a self-censoring effect. Since, we embedded our experiment in an already existing panel study, we were not able to increase the sample size.

To remedy this problem, we turn to another alternative and employ the modeling approach introduced by Eady (2017), which jointly models the indirect and direct items in the prediction of misreporting (i.e. self-censorship). By leveraging the information included in the direct item, this approach increases the statistical efficiency. In addition, we are able to include several variables that are predictive of self-censorship and thus help us to further reduce variance: the socio-demographics gender, age, and education as well as ideological leaning, political interest, and whether respondents feel free to discuss politics.

Table 5 shows the result for the misreporting equations separately for the two list experiments on free expression. In the US sample, self-censorship is driven by age and ideology but not in consistent ways across the two experiments. Importantly, priming hate speech law did not induce a greater propensity to misreport preferences for the free expression of offensive speech. In the German sample, we find that females are more likely to self-censor than males. Those with a college degree are less likely to misreport, but only in list experiment 2a. The effects of the political variables are again inconsistent. However, priming hate speech legislation has statistically significant effect on respondent's self-censoring behavior: $\beta = -1.50, p < .05$ in list experiment 2a and $\beta = -1.16, p < .01$ in list experiment 2b. Echoing the results of the previous sections, German respondents are less likely to self-censor their opinion that offensive speech should be restricted, when reminded of such legal norms.

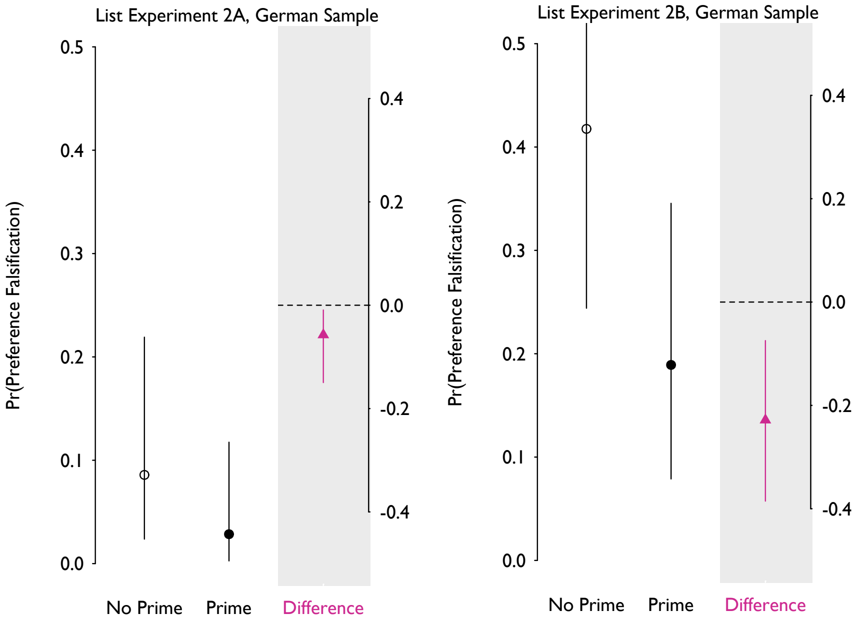
Figure 5 presents predicted probabilities and first differences to get a better sense of the substantive size of these effects. In list experiment 2A, German respondents who were not primed with hate speech law misreport their preference on free expression with a probability of 8 [95% CI: 2, 20] percent (fixing all covariates to their means). For those who received the hate speech prime, the probability of self-censorship reduces to a mere 3 [0, 12] percent. This amounts to an effect of -5 [-14, -1] percentage

Table 5: Regression Models of Self-Censorship

	Double List 1: Muslims Out		Double List 2: Offensive Opinion	
	Est.	SE	Est.	SE
USA				
Hate Speech Law	x	x	-0.40	0.36
Female	x	x	-0.38	0.41
Age/10	x	x	1.24	0.36
College	x	x	0.72	0.61
Ideology	x	x	-2.19	0.55
Political Interest	x	x	0.49	0.62
Feel free to discuss	x	x	0.93	0.39
(Intercept)	x	x	-0.87	0.78
Germany				
Hate Speech Law	3.29	(2.02)	-1.91	(0.74)
Female	-2.05	(1.76)	1.67	(0.69)
Age/10	-2.72	(1.68)	-0.18	(0.20)
College	0.81	(1.18)	-0.73	(0.84)
Ideology	-1.33	(0.70)	0.25	(0.18)
Political Interest	-0.49	(0.46)	0.57	(0.39)
Feel free to discuss	-0.85	(0.78)	-1.25	(0.43)
(Intercept)	14.24	(9.29)	-2.26	(1.82)

points. Results are stronger for list experiment 2B, where in the no-prime group the probability of self-censorship is 41 percent [24, 59] and 19 percent [8, 36]. The effect of priming hate speech law on self-censorship is thus -22 [-6, 38] percentage points. In sum, and counter to our expectations based on critics of hate speech legislation, restricting hate speech actually reduces self-censorship. What is driving this effect is the fact that more people are in favor of restricting offensive speech than would be willing to publicly admit. Hate speech law encourages these respondents to come forth with their preference for limiting free expression.

Figure 5: Predicted Probability of Self-Censorship



Discussion and Conclusion

The phenomenon of self-censorship has received considerable attention in the study of authoritarian regimes, where citizens are reluctant to voice their true opinion for fear of repression (Kuran, 1991, 1997; Gueorguiev et al., 2017; Jiang and Yang, 2016; Robinson and Tannenberg, 2018). Yet, neither is citizens' self-censoring behavior exclusively motivated by fear of violent state repression nor is it restricted to authoritarian contexts. Laws that restrict hateful or offensive speech targeted at protected groups are the subject of controversial debate around the world.

Our study is among the first to experimentally test two key claims about the likely unintended behavioral consequences of such laws. The first hypothesis states that citizens will self-censor offensive statements about protected minority groups such as Muslims. The second hypothesis posits a chilling effect, where citizens falsify their true policy preference. While the experimental results do not support our initial expectations, they do provide important new insights into the norms of free speech and the effects of restricting hate speech. While hate speech laws are unlikely to affect discriminatory statements, they have the potential to sway public opinion in favor of a more restrictive approach to free speech, both off- and online.

We do not deny the possibility of a priori positive consequences of hate speech legislation as often put forward by proponents of hate speech regulation. Norms implied by regulatory measures do signal societal norms and can have measurable feedback effects on related political attitudes, as has been shown, for instance, in the cases of same-sex marriage policies (Abou-Chadi and Finnigan, 2019), smoking bans (Pacheco, 2013), or welfare reform (Soss and Schram, 2007). In our context, one could therefore hypothesize that hate speech laws induce a true change in discriminatory attitudes. Our experimental setup does not allow us to speak to this idea. However, we argue that in the case of hate speech, regulatory efforts do not change political realities per se, which is why we do not regard attitudinal changes, such as a reduction in racism or derogatory views against people of a particular ethnicity, religion, sexual orientation, or gender as a particularly obvious outcome of hate speech regulation. Obviously, more research is needed to explore these relationships.

References

Abou-Chadi, Tarik and Ryan Finnigan. 2019. "Rights for Same-Sex Couples and Public Attitudes Toward Gays and Lesbians in Europe." *Comparative Political Studies* 52(6):868–895.

Aronow, Peter M, Alexander Coppock, Forrest W Crawford and Donald P Green. 2015. "Combining list experiment and direct question estimates of sensitive behavior prevalence." *Journal of Survey Statistics and Methodology* 3(1):43–66.

Berinsky, Adam J., Gregory A. Huber and Gabriel S. Lenz. 2012. "Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk." *Political Analysis* 20(3):351–368.

Blair, Graeme, Alexander Coppock and Margaret Moor. 2018. When to Worry About Sensitivity Bias: Evidence from 30 Years of List Experiments. Technical report Working Paper.

Blair, Graeme and Kosuke Imai. 2012. "Statistical Analysis of List Experiments." *Political Analysis* 20(1):47–77.

Bleich, Erik. 2011*a*. *The Freedom to Be Racist?: How the United States and Europe Struggle to Preserve Freedom and Combat Racism*. Oxford: Oxford University Press.

Bleich, Erik. 2011*b*. "The Rise of Hate Speech and Hate Crime Laws in Liberal Democracies." *Journal of Ethnic and Migration Studies* 37(6):917–934.

Brown, Alex. 2015. *Hate Speech Law: A Philosophical Examination*. Routledge.

Chong, Dennis. 1993. "How people think, reason, and feel about rights and liberties." *American journal of political science* pp. 867–899.

Dworkin, Ronald. 1999. *Freedom's law: the moral reading of the American Constitution*. OUP Oxford.

Eady, Gregory. 2017. "The Statistical Analysis of Misreporting on Sensitive Survey Questions." *Political Analysis* 25(2):241–259.

Egorov, Georgy, Sergei Guriev and Konstantin Sonin. 2009. "Why Resource-poor Dictators Allow Freer Media: A Theory and Evidence from Panel Data." *American Political Science Review* 103(4):645–668.

Fish, Stanley. 1994. *There's no such thing as free speech: And it's a good thing, too*. Oxford University Press.

Gibson, James L. 2006. "Enigmas of intolerance: Fifty years after Stouffer's communism, conformity, and civil liberties." *Perspectives on Politics* 4(1):21–34.

Gilens, Martin, Paul M Sniderman and James H Kuklinski. 1998. "Affirmative action and the politics of realignment." *British Journal of Political Science* 28(1):159–183.

Glynn, Adam N. 2013. "What Can We Learn with Statistical Truth Serum?: Design and Analysis of the List Experiment." *Public Opinion Quarterly* 77(S1):159–172.

Gueorguiev, Dimitar, Li Shao and Charles Crabtree. 2017. "Blurring the Lines: Rethinking Censorship Under Autocracy." *SSRN Electronic Journal* .

Hare, Ivan and James Weinstein. 2009. *Extreme Speech and Democracy*. Oxford: Oxford University Press.

Herz, Michael and Peter Molnar. 2012. *The Content and Context of Hate Speech: Rethinking Regulation and Responses*. Cambridge: Cambridge University Press.

Jiang, Junyan and Dali L. Yang. 2016. "Lying or Believing? Measuring Preference Falsification From a Political Purge in China." *Comparative Political Studies* 49(5):600–634.

King, Gary, Jennifer Pan and Margaret E. Roberts. 2013. "How Censorship in China Allows Government Criticism but Silences Collective Expression." *American Political Science Review* 107(2 (May)):1–18.

King, Gary, Jennifer Pan and Margaret E. Roberts. 2014. "Reverse-engineering censorship in China: Randomized experimentation and participant observation." *Science* 345(6199).

Kuran, Timur. 1991. "Now Out of Never: The Element of Surprise in the East European Revolution of 1989." *World Politics* 44(1):7–48.

Kuran, Timur. 1997. *Private truths, public lies: The social consequences of preference falsification*. Harvard University Press.

Loury, Glenn C. 1994. "Self-Censorship in Public Discourse: A Theory of "Political Correctness" and Related Phenomena." *Rationality and Society* 6(4):428–461.

Matsuda, Mari J. 1989. "Public response to racist speech: Considering the victim's story." *Michigan Law Review* 87(8):2320–2381.

Meserve, Stephen A. and Daniel Pemstein. 2018. "Google Politics: The Political Determinants of Internet Censorship in Democracies." *Political Science Research and Methods* 6(2):245–263.

Mill, John Stuart. 2011. *On Liberty*. Cambridge Library Collection - Philosophy Cambridge University Press.

Norwood v Director of Public Prosecutions. 2003. “[2003] EWHC 1564 (Admin) (03 July 2003).” https://www.bailii.org/ew/cases/EWHC/Admin/2003/1564.html.

Pacheco, Julianna. 2013. “Attitudinal Policy Feedback and Public Opinion: The Impact of Smoking Bans on Attitudes towards Smokers, Secondhand Smoke, and Antismoking Policies.” *Public Opinion Quarterly* 77(3):714–734.

Parekh, Bhikhu. 2012. *Is There a Case for Banning Hate Speech?* Cambridge: Cambridge University Press pp. 37–56.

Pew Research Center. 2016. “Europeans Fear Wave of Refugees Will Mean More Terrorism, Fewer Jobs.” https://www.pewresearch.org/global/wp-content/uploads/sites/2/2016/07/Pew-Research-Center-EU-Refugees-and-National-Identity-Report-FINAL-July-11-2016.pdf.

Pew Research Center. 2019. “What Americans Know About Religion.” https://www.pewforum.org/wp-content/uploads/sites/7/2019/07/Religious-Knowledge-full-draft-FOR-WEB-2.pdf.

Post, Robert. 2009. “Hate speech.” *Hare and Weinstein, Extreme Speech and Democracy* pp. 123–38.

Robinson, Darrel and Marcus Tannenberg. 2018. “Self-Censorship in Authoritarian States: Response Bias in Measures of Popular Support in China.” *SSRN Electronic Journal* .

Rosenfeld, Michel. 2002. “Hate speech in constitutional jurisprudence: a comparative analysis.” *Cardozo L. Rev.* 24:1523.

Soss, Joe and Sanford F. Schram. 2007. “A Public Transformed? Welfare Reform as Policy Feedback.” *American Political Science Review* 101(1):111–127.

Stanig, Piero. 2015. “Regulation of Speech and Media Coverage of Corruption: An Empirical Analysis of the Mexican Press.” *American Journal of Political Science* 59(1):175–193.

Strossen, Nadine. 2018. *HATE: Why We Should Resist it With Free Speech, Not Censorship*. Oxford: Oxford University Press.

Tourangeau, Roger and Ting Yan. 2007. “Sensitive questions in surveys.” *Psychological bulletin* 133(5):859.

Waldron, Jeremy. 2012. *The harm in hate speech*. Harvard University Press.

Wike, Richard and Katie Simmons. 2015. “Global support for principle of free expression, but opposition to some forms of speech.” *Pew Research Center* 18.

The Effect of Hate Speech Regulation on Self-Censorship

Online Appendix

Contents

Appendices 1

Appendix A Study design 2

A.1 Sampling and participants 2

A.2 Survey waves and fielding of experiments 2

Appendix B Deviations with respect to pre-analysis plan 4

Appendix C Supporting tables and figures 5

C.1 Balance tests 5

C.2 Descriptive results of list experiments 7

C.3 Descriptive results of direct items 8

C.4 Testing for design effects in list experiments 9

C.5 Full model results 11

C.6 Alternative identification strategy using the longitudinal list setup 13

C.7 Manipulation check results 17

C.8 Further robustness checks 20

Appendix D Original question wordings 22

Appendix E Software statement 37

Appendix A Study design

A.1 Sampling and participants

In partnership with the online survey firm YouGov, we initially recruited a total of 1,551 respondents from the US and 1,500 respondents from the German Pulse panel, a subset of YouGov's traditional survey panels in which members opt in to install passive metering software on their desktop and mobile devices. Respondents agreed to join a "Politics and Media" study with multiple survey waves. Their participation was rewarded using YouGov's proprietary points system and included a bonus for completing all waves in order to disincentivize attrition. Participation was voluntary and respondents were able to opt out from the passive metering part of the study at any time. Respondents were sampled according to YouGov's demographic/political targets then reweighted in order to obtain a sample that is representative of the U.S. population and the German online population, respectively (Rivers, 2006). Specifically, respondents were weighted to a sampling frame constructed from the full 2016 American Community Survey (ACS) 1-year sample (US sample) and the Best for Planning Survey sample (German sample). The sample cases were weighted to the sampling frame using propensity scores.

Collecting passive behavioral data that can be linked to survey responses poses privacy and ethical challenges beyond those typically associated with traditional social science research methods (Stier et al., 2019). While we did not make use of the passing tracking data for this study, we followed a strict protocol informed by IRB guidance and emerging best practices ?. This study was approved by the Institutional Review Boards of Princeton University (protocols 8327, 10014, and 10041) and the University of Southern California (UP-17-00513) and authorized by the University of Illinois via a designated IRB agreement. Before joining this specific study, respondents additionally agreed to a separate consent statement informing them, "Your participation is voluntary. Participation involves completion of a short survey and voluntary tracking of online media consumption. You may choose not to answer any or all questions. Furthermore, you are free to opt out of web tracking, which you may have previously agreed to participate in as part of the YouGov Pulse panel, at any time."

A.2 Survey waves and fielding of experiments

We conducted multiple survey waves. In **Germany**, the panel originally launched in July 2017 and was refreshed after a field time pause between Wave 5 and 6 with $1,023$ new respondents. The experiment was fielded in the Wave 7 (December 6–21, 2018; $N = 1,429$).

In the **United States**, a baseline survey was fielded between July 3–22, 2018 ($N = 1,551$) and several re-contact waves followed. The list experiment was originally embedded in Wave 5 (December 20, 2018–January 7, 2019; $N = 1,195$). However, after data collection we noticed that the survey provider had fielded a flawed version of the list experiments, due to which none of the respondents received the treatment items in any of the lists. As a result, the survey was fielded again to the same respondents (January 24, 2019–February 5, 2019; $N = 1,324$). Unfortunately, in this version the prime experiment was not properly implemented and all respondents received hate speech law prime regardless of their actual treatment status (only the question about the approval of these fictitious law contents was shown or hidden depending on the treatment status). As a consequence, the experiment had to be fielded again, now to a fresh sample of respondents from the Pulse panel (June 12, 2020–June 29, 2020; $N = 1,506$), which we refer to as Wave 9 (note again that there is no sample overlap

Table A1: Overview of experimental implementation across different waves

Wave	List experiment	Law prime	Prime treatment group
5	Failed (no sensitive items shown)	Properly implemented	W5, control W5, treatment
6	Properly implemented	Failed (everyone shown prime)	W6, treatment
9	Properly implemented	Properly implemented	W9, control W9, treatment

between this wave and any of the previous waves though). Table A1 in the Online Appendix provides an overview of the experimental implementation across the different waves.

Appendix B Deviations with respect to pre-analysis plan

Although the results we present here are based on the pre-analysis plan we registered prior to obtaining the data (see details at https://osf.io/fswm2), there were a few areas in which we had to deviate from the plan. In all cases, these deviations were due to errors in the fielding of the questionnaire by the survey provider or omissions in our pre-analysis plan. For the sake of transparency, we report each of these changes here:

-

Appendix C Supporting tables and figures

C.1 Balance tests

Table C1: Balance tests for hate speech legislation prime, US sample

	No prime	Prime	p-value
Female	0.53	0.56	0.09
Age	55.74	55.08	0.25
College	0.79	0.81	0.17
White	0.74	0.74	0.72
Ideology	2.96	2.90	0.22
Political interest	3.95	3.94	0.84
List experiment 1a	1.64	1.65	0.67
List experiment 1b	1.46	1.49	0.27
List experiment 2a	2.29	2.33	0.29
List experiment 2b	1.74	1.73	0.89

Table C2: Balance tests for list experiment 1, US sample

	List control	List treatment	p-value
Female	0.52	0.56	0.09
Age	55.61	55.59	0.98
College	0.79	0.80	0.30
White	0.74	0.74	1.00
Ideology	2.96	2.90	0.21
Political interest	3.95	3.96	0.77

Table C3: Balance tests for list experiment 2, US sample

	List control	List treatment	p-value
Female	0.54	0.54	0.79
Age	55.14	56.07	0.11
College	0.79	0.80	0.57
White	0.74	0.73	0.62
Ideology	2.95	2.90	0.33
Political interest	3.94	3.97	0.54

Table C4: Balance tests for hate speech legislation prime, German sample

	No prime	Prime	p-value
Female	0.44	0.45	0.77
Age	50.34	50.22	0.87
College	0.25	0.26	0.92
Ideology	4.60	4.54	0.61
Political interest	3.66	3.67	0.89
List experiment 1a	1.50	1.53	0.50
List experiment 1b	1.49	1.62	0.01
List experiment 2a	2.44	2.53	0.08
List experiment 2b	1.57	1.57	0.94

Table C5: Balance tests for list experiment 1, German sample

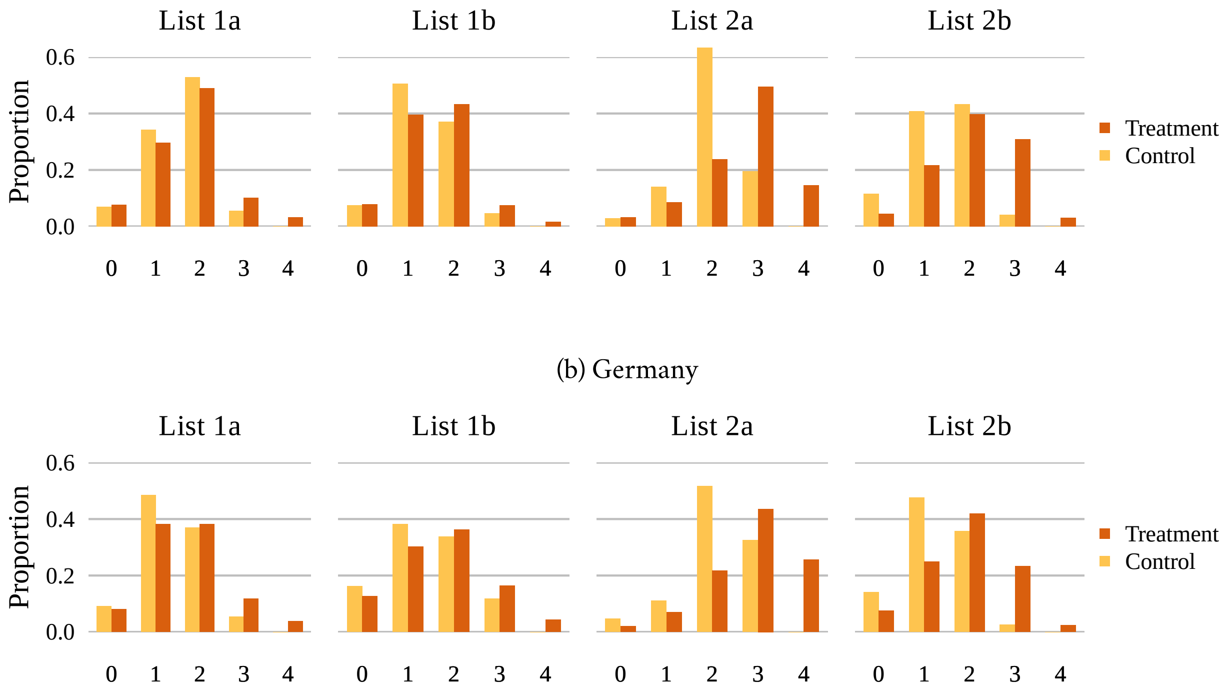
	List control	List treatment	p-value
Female	0.45	0.44	0.76
Age	50.61	49.94	0.36
College	0.25	0.26	0.57
Ideology	4.45	4.69	0.05
Political interest	3.66	3.66	0.95

Table C6: Balance tests for list experiment 2, German sample

	List control	List treatment	p-value
Female	0.44	0.45	0.87
Age	50.41	50.14	0.71
College	0.25	0.26	0.66
Ideology	4.48	4.65	0.18
Political interest	3.64	3.69	0.40

C.2 Descriptive results of list experiments

Figure C1: Reported counts in list experiments, by country and list treatment status.



(a) USA

(b) Germany

Table C7: Difference in reported counts on control lists, Wave 6 - Wave 5. Row proportions reported.

	-3	-2	-1	0	1	2	3
List 1a	0.00	0.02	0.09	0.50	0.35	0.04	
List 1b		0.01	0.13	0.69	0.13	0.02	0.00
List 2a	0.01	0.03	0.22	0.54	0.15	0.05	0.00
List 2b	0.00	0.01	0.14	0.68	0.15	0.02	0.00

C.3 Descriptive results of direct items

Figure C2: Support rates on direct items, by country

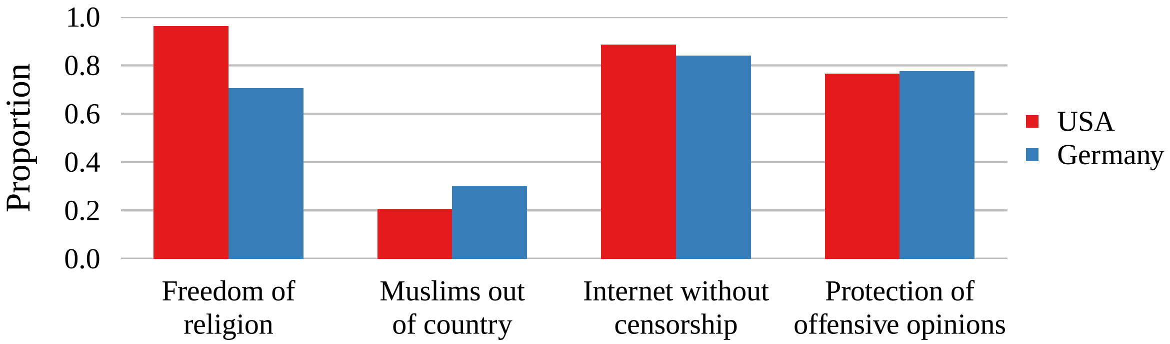


Table C8: Regression models of direct item support on hate speech law support

	Freedom of religion	Muslims out of country	Internet without censorship	Protection of offensive opinions
Support of hate speech law	−0.00	−0.02**	−0.05***	−0.09***
	(0.00)	(0.01)	(0.01)	(0.01)
R^2	0.00	0.01	0.04	0.08
Adj. R^2	−0.00	0.01	0.04	0.08
Num. obs.	1412	1405	1410	1410

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table C9: Difference in open support on various items, Wave 6 - Wave 5 (among respondents primed in both waves).

	-1	0	1
Freedom of religion	0.01	0.97	0.01
Muslims out of country	0.06	0.91	0.04
Internet without censorship	0.08	0.86	0.06
Protection of offensive opinions	0.09	0.81	0.10

C.4 Testing for design effects in list experiments

In this section we report statistical tests for the null hypothesis of 'no design effect' in our list experiments. A design effect occurs when responses to the non-sensitive control items change depending on the inclusion of the sensitive item (Blair and Imai, 2012). We fail to reject the null of no design effect in all list experiments and for both the US and the German sample.

Table C10: Tests for no design effects in list experiments, US sample

Outcome/treatment status	est.	s.e.
List experiment 1a		
$p_i(Y_i(0) = 0, Z_i = 1)$	-0.01	0.01
$p_i(Y_i(0) = 1, Z_i = 1)$	0.04	0.02
$p_i(Y_i(0) = 2, Z_i = 1)$	0.08	0.01
$p_i(Y_i(0) = 3, Z_i = 1)$	0.03	0.00
$p_i(Y_i(0) = 0, Z_i = 0)$	0.08	0.01
$p_i(Y_i(0) = 1, Z_i = 0)$	0.30	0.01
$p_i(Y_i(0) = 2, Z_i = 0)$	0.45	0.02
$p_i(Y_i(0) = 3, Z_i = 0)$	0.02	0.01
Bonferroni-corr. p-value	0.49	
List experiment 1b		
$p_i(Y_i(0) = 0, Z_i = 1)$	0.00	0.01
$p_i(Y_i(0) = 1, Z_i = 1)$	0.10	0.02
$p_i(Y_i(0) = 2, Z_i = 1)$	0.04	0.01
$p_i(Y_i(0) = 3, Z_i = 1)$	0.02	0.00
$p_i(Y_i(0) = 0, Z_i = 0)$	0.08	0.01
$p_i(Y_i(0) = 1, Z_i = 0)$	0.40	0.02
$p_i(Y_i(0) = 2, Z_i = 0)$	0.33	0.02
$p_i(Y_i(0) = 3, Z_i = 0)$	0.03	0.01
Bonferroni-corr. p-value	0.63	
List experiment 2a		
$p_i(Y_i(0) = 0, Z_i = 1)$	0.00	0.01
$p_i(Y_i(0) = 1, Z_i = 1)$	0.05	0.01
$p_i(Y_i(0) = 2, Z_i = 1)$	0.45	0.02
$p_i(Y_i(0) = 3, Z_i = 1)$	0.15	0.01
$p_i(Y_i(0) = 0, Z_i = 0)$	0.03	0.00
$p_i(Y_i(0) = 1, Z_i = 0)$	0.09	0.01
$p_i(Y_i(0) = 2, Z_i = 0)$	0.19	0.02
$p_i(Y_i(0) = 3, Z_i = 0)$	0.05	0.01
Bonferroni-corr. p-value	0.61	
List experiment 2b		
$p_i(Y_i(0) = 0, Z_i = 1)$	0.07	0.01
$p_i(Y_i(0) = 1, Z_i = 1)$	0.26	0.02
$p_i(Y_i(0) = 2, Z_i = 1)$	0.30	0.01
$p_i(Y_i(0) = 3, Z_i = 1)$	0.03	0.00
$p_i(Y_i(0) = 0, Z_i = 0)$	0.05	0.01
$p_i(Y_i(0) = 1, Z_i = 0)$	0.14	0.01
$p_i(Y_i(0) = 2, Z_i = 0)$	0.13	0.02
$p_i(Y_i(0) = 3, Z_i = 0)$	0.01	0.01
Bonferroni-corr. p-value	1	

Table C11: Tests for no design effects in list experiments (German sample)

Outcome/treatment status	est.	s.e.
List experiment 1a		
$p_i(Y_i(0) = 0, Z_i = 1)$	0.01	0.01
$p_i(Y_i(0) = 1, Z_i = 1)$	0.11	0.03
$p_i(Y_i(0) = 2, Z_i = 1)$	0.10	0.02
$p_i(Y_i(0) = 3, Z_i = 1)$	0.04	0.01
$p_i(Y_i(0) = 0, Z_i = 0)$	0.08	0.01
$p_i(Y_i(0) = 1, Z_i = 0)$	0.37	0.02
$p_i(Y_i(0) = 2, Z_i = 0)$	0.27	0.02
$p_i(Y_i(0) = 3, Z_i = 0)$	0.02	0.01
Bonferroni-corr. p-value	1	
List experiment 1b		
$p_i(Y_i(0) = 0, Z_i = 1)$	0.03	0.02
$p_i(Y_i(0) = 1, Z_i = 1)$	0.11	0.03
$p_i(Y_i(0) = 2, Z_i = 1)$	0.09	0.02
$p_i(Y_i(0) = 3, Z_i = 1)$	0.04	0.01
$p_i(Y_i(0) = 0, Z_i = 0)$	0.13	0.01
$p_i(Y_i(0) = 1, Z_i = 0)$	0.27	0.02
$p_i(Y_i(0) = 2, Z_i = 0)$	0.25	0.02
$p_i(Y_i(0) = 3, Z_i = 0)$	0.07	0.01
Bonferroni-corr. p-value	1	
List experiment 2a		
$p_i(Y_i(0) = 0, Z_i = 1)$	0.03	0.01
$p_i(Y_i(0) = 1, Z_i = 1)$	0.07	0.02
$p_i(Y_i(0) = 2, Z_i = 1)$	0.37	0.02
$p_i(Y_i(0) = 3, Z_i = 1)$	0.26	0.02
$p_i(Y_i(0) = 0, Z_i = 0)$	0.02	0.01
$p_i(Y_i(0) = 1, Z_i = 0)$	0.04	0.01
$p_i(Y_i(0) = 2, Z_i = 0)$	0.15	0.02
$p_i(Y_i(0) = 3, Z_i = 0)$	0.07	0.02
Bonferroni-corr. p-value	1	
List experiment 2b		
$p_i(Y_i(0) = 0, Z_i = 1)$	0.07	0.02
$p_i(Y_i(0) = 1, Z_i = 1)$	0.29	0.03
$p_i(Y_i(0) = 2, Z_i = 1)$	0.23	0.02
$p_i(Y_i(0) = 3, Z_i = 1)$	0.02	0.01
$p_i(Y_i(0) = 0, Z_i = 0)$	0.07	0.01
$p_i(Y_i(0) = 1, Z_i = 0)$	0.18	0.02
$p_i(Y_i(0) = 2, Z_i = 0)$	0.13	0.02
$p_i(Y_i(0) = 3, Z_i = 0)$	0.00	0.01
Bonferroni-corr. p-value	1	

C.5 Full model results

Table C12: Effect of priming hate speech legislation on direct support rates of various statements (US sample)

	No prime		Prime		Effect	
	Est.	SE	Est.	SE	Est.	SE
People should be able to practice their religion freely in our country.	0.97	0.01	0.97	0.00	0.01	0.01
Muslims out of USA. Protect the American People!	0.18	0.02	0.18	0.01	-0.01	0.02
It is important that people can use the Internet without government censorship.	0.93	0.01	0.87	0.01	-0.06	0.02
People should be allowed to express unpopular opinions in public, even those that are deeply offensive to other people.	0.75	0.02	0.79	0.01	0.04	0.02

Table C13: Effect of priming hate speech legislation on direct support rates of various statements (German sample)

	No prime		Prime		Effect	
	Est.	SE	Est.	SE	Est.	SE
People should be able to practice their religion freely in our country.	0.72	0.02	0.75	0.02	0.02	0.03
Muslims out of Germany. Protect the German People!	0.31	0.02	0.25	0.02	-0.06	0.03
It is important that people can use the Internet without government censorship.	0.89	0.01	0.80	0.02	-0.09	0.02
People should be allowed to express unpopular opinions in public, even those that are deeply offensive to other people.	0.81	0.02	0.75	0.02	-0.06	0.03

Table C14: Effect of priming hate speech legislation on self-censorship (German sample)

	No prime		Prime		Effect	
	Est.	SE	Est.	SE	Est.	SE
Muslims out of Germany. Protect the German People!	-0.06	0.05	-0.01	0.06	0.05	0.08
People should be allowed to express unpopular opinions in public, even those that are deeply offensive to other people.	-0.12	0.05	-0.08	0.05	0.04	0.07

Table C15: Effect of priming hate speech legislation on self-censorship (US sample)

	No prime		Prime		Effect	
	Est.	SE	Est.	SE	Est.	SE
Muslims out of USA Protect the American People!	-0.05	0.04	-0.01	0.03	0.04	0.05
People should be allowed to express unpopular opinions in public, even those that are deeply offensive to other people.	-0.06	0.05	-0.10	0.03	-0.04	0.06

C.6 Alternative identification strategy using the longitudinal list setup

The fact that respondents received at least once both the control and the treatment version of each list across Waves 5 and 6 allows us to implement an alternative identification strategy for the effects of the prime on self-censorship.

Table C16: Regression models of direct support rates of anti-Muslim sentiment

	Model 1	Model 2	Model 3
Hate speech law prime	−0.01	0.01	0.04
	(0.02)	(0.03)	(0.03)
Sensitive item support (list)		0.16^{***}	0.13^{***}
		(0.03)	(0.03)
Prime X Sensitive item support (list)		−0.05	−0.09
		(0.05)	(0.05)
Female			0.02
			(0.02)
Age/10			0.00
			(0.01)
College			-0.13^{***}
			(0.03)
Ideology			0.11^{***}
			(0.01)
Political interest			−0.00
			(0.01)
Feel free to discuss			0.06^{***}
			(0.02)
R^2	0.00	0.03	0.20
Adj. R^2	−0.00	0.02	0.19
Num. obs.	1155	1083	961

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table C17: Regression models of direct support rates of free speech for unpopular and offensive opinions

	Model 1	Model 2	Model 3
Hate speech law prime	−0.02	−0.02	−0.02
	(0.02)	(0.04)	(0.04)
Sensitive item support (list)		0.16^{***}	0.15^{***}
		(0.04)	(0.04)
Prime X Sensitive item support (list)		0.01	0.00
		(0.05)	(0.05)
Female			-0.09^{***}
			(0.03)
Age/10			0.00
			(0.01)
College			0.01
			(0.04)
Ideology			0.02^{*}
			(0.01)
Political interest			0.02
			(0.01)
Feel free to discuss			−0.00
			(0.02)
R^2	0.00	0.04	0.06
Adj. R^2	−0.00	0.04	0.05
Num. obs.	1161	1088	963

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table C18: Regression models of self-censorship on anti-Muslim opinion, longitudinal list-based measure

	Full sample		Anti-Muslim opinions only	
	Model 1	Model 2	Model 3	Model 4
Hate speech law prime	0.00	0.01	0.04	0.05
	(0.03)	(0.03)	(0.04)	(0.04)
Female		0.02		−0.00
		(0.03)		(0.04)
Age/10		−0.01		−0.01
		(0.01)		(0.02)
College		−0.05		0.09
		(0.05)		(0.05)
Ideology		-0.03^{*}		-0.15^{***}
		(0.01)		(0.02)
Political interest		−0.00		−0.02
		(0.02)		(0.02)
Feel free to discuss		0.01		-0.06^{*}
		(0.02)		(0.03)
R^2	0.00	0.01	0.00	0.25
Adj. R^2	−0.00	0.00	−0.00	0.23
Num. obs.	983	872	408	366

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table C19: Regression models of self-censorship on free speech scepticism, longitudinal list-based measure

	Full sample		Free speech sceptics only	
	Model 1	Model 2	Model 3	Model 4
Hate speech law prime	−0.02	−0.01	−0.02	−0.02
	(0.03)	(0.03)	(0.05)	(0.05)
Female		0.05		−0.10*
		(0.03)		(0.05)
Age/10		0.02*		0.01
		(0.01)		(0.02)
College		−0.09*		−0.03
		(0.04)		(0.06)
Ideology		0.01		0.04*
		(0.01)		(0.02)
Political interest		−0.01		0.04
		(0.01)		(0.02)
Feel free to discuss		0.03		−0.00
		(0.02)		(0.03)
R^2	0.00	0.02	0.00	0.04
Adj. R^2	−0.00	0.02	−0.00	0.02
Num. obs.	989	879	394	345

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

C.7 Manipulation check results

To investigate whether the hate speech law prime delivered the information it was intended to deliver, we implemented two manipulation checks in US Wave 9 of our study. Directly following the direct items, respondents from both the hate speech law prime treatment and control groups were asked:

MANIPULATION CHECK 1, US SURVEY, WAVE 9

If you were to speak up and **give your views on Muslims in the US** during an online discussion, how concerned would you be that the following would occur:

- Others would criticize my views as offensive.
- Someone would post critical comments about my views on social media.
- Someone would file a complaint claiming that my views violated online community standards.
- My employer might discipline me for my views.
- The police would investigate me for expressing my views.
- I would face legal prosecution for expressing my views.

1. Not at all concerned
2. Slightly concerned
3. Somewhat concerned
4. Moderately concerned
5. Extremely concerned
6. Don’t know

MANIPULATION CHECK 2, US SURVEY, WAVE 9

If you were to speak up and **give your views on free speech** during an online discussion, how concerned would you be that the following would occur:

- Others would criticize my views as offensive.
- Someone would post critical comments about my views on social media.
- Someone would file a complaint claiming that my views violated online community standards.
- My employer might discipline me for my views.
- The police would investigate me for expressing my views.
- I would face legal prosecution for expressing my views.

1. Not at all concerned
2. Slightly concerned
3. Somewhat concerned
4. Moderately concerned
5. Extremely concerned
6. Don't know

Tables C20 and C21 report computed means on the five-point scales for both the no-prime and the prime group as well as the results from a t-test on the difference of means.

In line with the content of the hate speech law, members of the prime group reported higher levels of concern about police investigation and legal prosecution as a consequence of expressing their views on Muslims in the US. The differences on these two items are somewhat weaker for the more abstract situation of giving views on free speech. Instead, members of the prime group reported higher levels of concern about their employer possibly disciplining them for their views. We find no meaningful differences on any of the other items, whereby in 9 out of 10 cases the prime group shows higher levels than the control group.

Table C20: T-tests for hate speech legislation prime manipulation check 1.

	No prime	Prime	p-value
Others would criticize my views as offensive.	2.22	2.25	0.73
Someone would post critical comments about my views on social media.	2.39	2.38	0.92
The police would investigate me for expressing my views.	1.88	2.04	0.06
I would face legal prosecution for expressing my views.	1.91	2.08	0.05
My employer might discipline me for my views.	2.01	2.13	0.21

Note: Question asked in Wave 9 following the direct items. Question text: "If you were to speak up and give your views on Muslims in the US during an online discussion, how concerned would you be that the following would occur." Answering scale ranging from 1 = Not at all concerned to 5 = Extremely concerned.

Table C21: T-tests for hate speech legislation prime manipulation check 2.

	No prime	Prime	p-value
Others would criticize my views as offensive.	2.05	2.08	0.73
Someone would post critical comments about my views on social media.	2.12	2.16	0.59
The police would investigate me for expressing my views.	1.80	1.90	0.22
I would face legal prosecution for expressing my views.	1.79	1.90	0.16
My employer might discipline me for my views.	1.86	2.03	0.05

Note: Question asked in Wave 9 following the direct items. Question text: "Next, if you were to speak up and give your views on free speech during an online discussion, how concerned would you be that the following would occur." Answering scale ranging from 1 = Not at all concerned to 5 = Extremely concerned.

C.8 Further robustness checks

Table C22: Effect of priming hate speech legislation once or twice on direct support rates of various statements (US sample)

	Primed once		Primed twice		Effect	
	Est.	SE	Est.	SE	Est.	SE
People should be able to practice their religion freely in our country.	0.97	0.01	0.98	0.01	0.01	0.01
Muslims out of USA. Protect the American People!	0.20	0.02	0.17	0.02	-0.03	0.03
It is important that people can use the Internet without government censorship.	0.87	0.02	0.85	0.02	-0.02	0.02
People should be allowed to express unpopular opinions in public, even those that are deeply offensive to other people.	0.87	0.02	0.79	0.02	-0.02	0.03

Table C23: Effect of priming hate speech legislation once or twice on self-censorship (US sample)

	Primed once		Primed twice		Effect	
	Est.	SE	Est.	SE	Est.	SE
Muslims out of USA Protect the American People!	0.01	0.05	-0.02	0.05	-0.03	0.07
People should be allowed to express unpopular opinions in public, even those that are deeply offensive to other people.	-0.09	0.05	-0.15	0.05	-0.06	0.07

Table C24: T-tests for effect of receiving list treatment item on direct item support for those not receiving hate speech law prime.

	Control lists	Treatment lists	p-value
People should be able to practice their religion freely in our country.	0.96	0.96	0.66
Muslims out of USA. Protect the American People!	0.23	0.20	0.30
It is important that people can use the Internet without government censorship.	0.88	0.92	0.00
People should be allowed to express unpopular opinions in public, even those that are deeply offensive to other people.	0.79	0.76	0.17

Note: Control lists refers to respondents in Wave 5, treatment lists to respondents in Wave 9.

Table C25: T-tests for effect of receiving list treatment item on direct item support for those receiving hate speech law prime.

	Control lists	Treatment lists	p-value
People should be able to practice their religion freely in our country.	0.97	0.96	0.44
Muslims out of USA. Protect the American People!	0.20	0.19	0.53
It is important that people can use the Internet without government censorship.	0.87	0.87	0.94
People should be allowed to express unpopular opinions in public, even those that are deeply offensive to other people.	0.78	0.74	0.12

Note: Control lists refers to respondents in Wave 5, treatment lists to respondents in Wave 9.

Table C26: Prevalence of sensitive list items in Wave 6 by hate speech law prime treatment status in Wave 5

		Est.	SE
List 1	Prevalence of sensitive item, list (prime)	0.15	0.04
	Prevalence of sensitive item, list (no prime)	0.21	0.04
	Difference (prime - no prime)	-0.06	0.06
List 2	Prevalence of sensitive item, list (prime)	0.66	0.04
	Prevalence of sensitive item, list (no prime)	0.68	0.04
	Difference (prime - no prime)	-0.02	0.06

Appendix D Original question wordings

Figure D1: Design of List 1A

LIST 1A, US SURVEY

In the following you will find statements posted on social media that some people support or oppose. After you have read all statements, please tell us HOW MANY of them you support. We do not want to know which ones, just HOW MANY.

- "Marijuana is harmless. Legalize it!"
- "Vaccines save lives. Mandatory vaccination for all children!"
- "Stop killing babies. Make Abortion illegal in ALL cases!"
- **[treatment item]** "Muslims out of USA. Protect the American People!"

○ 0 ○ 1 ○ 2 ○ 3 ○ [4]

LIST 1A, GERMAN SURVEY

Im Folgenden sehen Sie einige Aussagen, die in den sozialen Medien gemacht wurden und die manche Leute unterstützen, andere ablehnen. Nachdem Sie alle Aussagen gelesen haben, sagen Sie uns bitte WIE VIELE davon sie unterstützen. Uns interessiert nicht welche, nur WIE VIELE.

- "Marihuana ist harmlos. Legalisierung jetzt!"
- "Impfungen retten Leben. Impfpflicht für alle Kinder!"
- "Stoppt das Töten von Babies. Macht Abtreibung in ALLEN Fällen illegal!"
- **[treatment item]** "Muslime raus aus Deutschland. Schützt das Deutsche Volk!"

○ 0 ○ 1 ○ 2 ○ 3 ○ [4]

HATE SPEECH EXPERIENCE

"Hate Speech" describes when someone is verbally attacked because of personal attributes, such as religion, ethnic origin, nationality, sex, or opinions. Please select all of the following that apply to you.

○ I have personally been verbally attacked with hate speech online.

○ I have experienced how others have been verbally attacked with hate speech online.

○ None of the above.

Figure D2: Design of List 1B

LIST 1B, US SURVEY

Here is another set of statements posted on social media that some people support or oppose. Please tell us again HOW MANY of them you support. We do not want to know which ones, just HOW MANY.

- "More Women in Tech. Affirmative Action now!"
- "Guns don't kill people. People kill people."
- "Save the planet. Raise the taxes on gasoline!"
- **[treatment item]** "Muslims out of USA. Protect the American People!"

○ 0 ○ 1 ○ 2 ○ 3 ○ [4]

LIST 1B, GERMAN SURVEY

Hier sind weitere Aussagen, die in den sozialen Medien gemacht wurden und die manche Leute unterstützen, andere ablehnen. Nachdem Sie alle Aussagen gelesen haben, sagen Sie uns bitte WIE VIELE davon sie unterstützen. Uns interessiert nicht welche, nur WIE VIELE.

- "Mehr Frauen in Technik-Berufen. Frauenquote sofort!"
- "Nicht Schusswaffen töten Menschen. Menschen töten Menschen!"
- "Rettet den Planeten. Höhere Steuern auf Benzin!"
- **[treatment item]** "Muslime raus aus Deutschland. Schützt das Deutsche Volk!"

○ 0 ○ 1 ○ 2 ○ 3 ○ [4]

HATE SPEECH REGULATION PREFERENCES

Would you support or oppose a law that would make it illegal to make insulting or hateful statements about...

- Germans? (American Citizens?)
- Muslims?
- Jews?
- Women?
- Christians?
- Neo Nazis?
- the Government?

○ Very much oppose

○ Rather oppose

○ Neither/nor

Figure D3: Caption

LIST 2A, US SURVEY

In the following you will find statements that some people support or oppose. After you have read all statements, please tell us HOW MANY of them you support. We do not want to know which ones, just HOW MANY.

- “People should be able to make statements that criticize the government publicly.”
- “Media organizations should be able to publish information about large political protests in our country.”
- “Government should be able to stop a news media outlet from publishing biased or inaccurate information.”
- **[treatment item]** “People should be allowed to express unpopular opinions in public, even those that are deeply offensive to other people.”

○ 0 ○ 1 ○ 2 ○ 3 ○ [4]

LIST 2A, GERMAN SURVEY

Im Folgenden sehen Sie einige Aussagen, die manche Leute unterstützen, andere ablehnen. Nachdem Sie alle Aussagen gelesen haben, sagen Sie uns bitte WIE VIELE davon sie unterstützen. Uns interessiert nicht welche, nur WIE VIELE.

- “Die Leute sollten die Regierung öffentlich kritisieren dürfen.”
- “Die Medien sollten über große politische Proteste im Land berichten dürfen.”
- “Die Regierung sollte die Medien davon abhalten dürfen, einseitige oder falsche Informationen zu veröffentlichen.”
- **[treatment item]** “Die Leute sollten unbeliebte Meinungen öffentlich äußern dürfen, selbst wenn andere diese Meinungen zutiefst anstößig finden.”

○ 0 ○ 1 ○ 2 ○ 3 ○ [4]

HATE SPEECH IDENTIFICATION

Which of the following would you label as hate speech?

- A person calling an ethnic minority a racial slur.
- A person calling a woman a vulgar name.
- A person who says that illegal immigrants should be deported.
- A person who says Germany/the USA is an evil country.
- A person who says Islam is taking over Europe/the USA.
- A person calling another person with conservative views a Nazi.

○ Hate speech

○ No hate speech

Figure D4: Design of List 2B

LIST 2B, US SURVEY

Here is another set of statements that some people support or oppose. Please tell us again HOW MANY of them you support. We do not want to know which ones, just HOW MANY.

- “Marijuana should be legalized, even if it may be harmful for some people.”
- “Vaccination should be mandatory for all children, even if parents oppose it.”
- “Abortion should be illegal, even if there is a health risk for the mother.”
- **[treatment item]** “People should be allowed to express unpopular opinions in public, even those that are deeply offensive to other people.”

○ 0 ○ 1 ○ 2 ○ 3 ○ [4]

LIST 2B, GERMAN SURVEY

Hier sind weitere Aussagen, die manche Leute unterstützen, andere ablehnen. Nachdem Sie alle Aussagen gelesen haben, sagen Sie uns bitte WIE VIELE davon sie unterstützen. Uns interessiert nicht welche, nur WIE VIELE.

- “Marijuana sollte legalisiert warden, auch wenn es für manche Leute schädlich ist.”
- “Impfungen sollte für alle Kinder verpflichtend sein, auch wenn die Eltern dagegen sind.”
- “Abtreibung sollte verboten sein, auch bei Gesundheitsrisiken für die Mutter.”
- **[treatment item]** “Die Leute sollten unbeliebte Meinungen öffentlich äußern dürfen, selbst wenn andere diese Meinungen zutiefst anstößig finden.”

○ 0 ○ 1 ○ 2 ○ 3 ○ [4]

RESPONSIBILITY FOR ACTION AGAINST HATE SPEECH

To what extent, if at all, do you think each of the following groups should take responsibility in taking steps against online hate speech?

- People who are victims of online hate speech
- Other users who witness the behavior
- Online services such as social media platforms or other websites
- Policymakers
- Law enforcement
- Employers of distributors of hate speech

○ No responsibility at all

○ Rather no responsibility

○ Some responsibility

Figure D5: Design of attention check before hate speech regulation prime

ATTENTION CHECK BEFORE PRIME, US SURVEY

People have different opinions about abolishing the switch to daylight saving time. Some would like to get rid of daylight saving time, others to get rid of standard time, others want everything to stay as it is. Specifically, we want to know whether you actually take your time to read the questions and follow our instructions. To demonstrate that you read this far, skip this question and just type „read“ in the text field below.

○ Very much oppose

○ Rather oppose

○ Neither oppose nor support

○ Rather support

○ Very much support

○ Other: ________________

ATTENTION CHECK BEFORE PRIME, GERMAN SURVEY

Leute haben unterschiedliche Meinungen zur Abschaffung der Zeitumstellung. Einige würden gerne die Sommerzeit abschaffen, andere die Winterzeit, andere hätten gerne, dass alles so bleibt, wie es ist. Wir möchten von Ihnen wissen, ob Sie sich eigentlich die Zeit nehmen die Fragen zu lesen und den Anweisungen zu folgen. Um zu zeigen, dass Sie bis hierhin gelesen haben, tragen Sie bitte “gelesen” in das Feld “Andere” unten ein.

○ Lehne voll und ganz ab

○ Lehne eher ab

○ Weder noch

○ Unterstütze eher

○ Unterstütze voll und ganz

○ Andere: ________________

FEELING TOWARDS DISCUSSING POLITICS

When you discuss politics with others, how free or unrestricted do you feel?

○ I don’t feel free to discuss it with anyone

○ I don’t feel free to discuss it with many people

○ I feel free to discuss it with a few

○ I feel free to discuss it with anyone

○ I never discuss politics with other people

Figure D6: Design of hate speech regulation prime, U.S. survey

HATE SPEECH REGULATION PRIME, US SURVEY

As you may have heard, the government is making serious efforts to combat online hate speech. This could mean that a large number of social media posts with offensive or hateful content will be deleted and legally prosecuted. The content of hate speech legislation that is currently discussed is described in the following text. Please read it very carefully and make sure you understand it. Please read it very carefully and make sure you understand it.

"A person is guilty of an offense if she sends a message over an online platform which

- *uses threatening, abusive or insulting words, or*
- *displays any writing, image or video which is threatening, abusive or insulting,*

if she intends thereby to stir up hatred against a religious group.
A person guilty of an offense under this law is liable for a prison term not exceeding six months or a fine or both.
This law does not prohibit or restrict discussion, criticism or expressions of antipathy, dislike, ridicule, insult or abuse of particular religions or the beliefs or practices of their adherents."

Having carefully read the content of this hate speech legislation, do you favor or oppose this law?

○ Very much oppose

○ Rather oppose

○ Neither oppose nor support

○ Rather support

○ Very much support

Figure D7: Design of hate speech regulation prime, German survey

HATE SPEECH REGULATION PRIME, GERMAN SURVEY

Wie Sie vielleicht gehört haben, bemüht sich die Regierung sehr ernsthaft Online-Hassrede zu bekämpfen. Das bedeutet, dass eine große Anzahl an Social-Media-Nachrichten mit beleidigenden oder hasserfüllten Inhalten gelöscht und strafrechtlich verfolgt werden.
Der Inhalt eines Hate-Speech-Gesetzes, das derzeit in der Diskussion ist, wird im folgenden Text beschrieben. Bitte lesen Sie den Text sehr sorgfältig und stellen Sie sicher, dass Sie ihn verstehen.

"Eine Person begeht eine Straftat, wenn sie eine Nachricht über eine Online-Plattform versendet, die

- *drohende, abwertende oder beleidigende Worte oder*
- *drohende, abwertende oder beleidigende Texte, Bilder oder Videos enthält*

und die Absicht hat, damit Hass gegen eine religiöse Gruppe zu schüren. Einer Person die sich im Sinne dieses Gesetzes strafbar macht, drohen eine Gefängnisstrafe von maximal sechs Monaten oder eine Geldstrafe oder beides.
Das Gesetz verbietet nicht die Diskussion und Kritik, den Ausdruck von Ablehnung oder das Lächerlichmachen, Beleidigen und Abwerten von bestimmten religiösen Glaubensinhalten oder -praktiken ihrer Anhänger."

Nachdem Sie den Inhalt des Hate-Speech-Gesetzes sorgfältig gelesen haben, unterstützen Sie das Gesetz oder lehnen Sie es ab?

○ Stimme überhaupt nicht zu

○ Stimme eher nicht zu

○ Teils/teils

○ Stimme eher zu

○ Stimme voll und ganz zu

○ Weiß nicht

Figure D8: Design of direct attitude measure

DIRECT ITEMS, US SURVEY

Here you can find several statements made on social media that some people support while others oppose. Do you support or oppose these statements?

- "People should be able to practice their religion freely in our country."
- "Muslims out of USA. Protect the American People!"
- "It is important that people can use the Internet without government censorship."
- "People should be allowed to express unpopular opinions in public, even those that are deeply offensive to other people."

○ Oppose

○ Support

DIRECT ITEMS, GERMAN SURVEY

Hier sind einige Aussagen, die in den sozialen Medien gemacht wurden und die manche Leute unterstützen, andere ablehnen. Unterstützen Sie diese Aussagen oder lehnen Sie sie ab?

- "Die Leute sollten ihre Religion in unserem Land frei ausüben dürfen."
- "Muslime raus aus Deutschland. Schützt das Deutsche Volk!"
- "Es ist wichtig, dass die Leute das Internet ohne Zensur durch die Regierung nutzen können."
- "Die Leute sollten unbeliebte Meinungen öffentlich äußern dürfen, selbst wenn andere diese Meinungen zutiefst anstößig finden."

○ Lehne ab

○ Unterstütze

POLITICAL INTEREST

How regularly do you follow politics?

○ Most of the time

○ Some of the time

○ Only now and then

○ Hardly at all

○ Don’t know

POLITICAL IDEOLOGY

In general, would you describe your political views as...

○ Very conservative

○ Conservative

○ Moderate

○ Liberal

○ Very liberal

○ Not sure

CONGRESS CONTROL PREFERENCE
Which party would you prefer to control Congress after the midterm elections?

○ Democrats

○ Republicans

○ Divided between House and Senate

○ None of the above

PRESIDENTIAL APPROVAL

Do you approve or disapprove of the way Donald Trump is handling his job as president?

○ Strongly approve

○ Somewhat approve

○ Neither approve nor disapprove

○ Somewhat disapprove

○ Strongly disapprove

SOCIAL MEDIA USE

Do you have accounts on any of the following social media services? (check all that apply):

- Twitter
- Facebook
- Instagram
- LinkedIn
- Snapchat
- WhatsApp
- Reddit

TWITTER USAGE FREQUENCY

In the last survey you told us that you have a Twitter account. Today we want to learn more about your Twitter use. How frequently do you:

1. Check Twitter
2. Post messages on Twitter

○ Almost constantly

○ Several times a day

○ About once a day

○ 3 to 6 days a week

○ 1 to 2 days a week

○ Every few weeks

○ Less often

○ Never

○ Don’t know

FACEBOOK USAGE FREQUENCY

In the last survey you told us that you have a Facebook account. Today we want to learn more about your Facebook use. How frequently do you:

1. Check Facebook
2. Post messages on Facebook

○ Almost constantly

○ Several times a day

○ About once a day

○ 3 to 6 days a week

○ 1 to 2 days a week

○ Every few weeks

○ Less often

○ Never

○ Don’t know

RACIAL RESENTMENT
Here you can find several statements with which some people agree while others do not. How about you? Please state your view on these issues.

1. Irish, Italians, Jewish and many other minorities overcame prejudice and worked their way up. Blacks should do the same without any special favors.
2. Generations of slavery and discrimination have created conditions that make it difficult for blacks to work their way out of the lower class
3. Over the past few years, blacks have gotten less than they deserve.
4. It's really a matter of some people not trying hard enough; if blacks would only try harder they could be just as well off as whites.

○ Strongly disagree

○ Somewhat disagree

○ Neither/nor

○ Somewhat agree

○ Strongly agree

○ Don't know

GENDER
What is your gender?

○ Male

○ Female

AGE
What is your age? OPEN ANSWER [years]

EDUCATION
Which of the following describes best your education?

○ No high school

○ High school graduate

○ Some college

○ 2-year

○ 4-year

○ Post-grad

RELIGION
What is your religion?

○ Protestant

○ Roman Catholic

○ Mormon

○ Eastern or Greek Orthodox

○ Jewish

○ Muslim

○ Buddhist

○ Hindu

○ Atheist

○ Agnostic

○ Nothing in particular

○ Something else

IMPORTANCE OF RELIGION
As how important do you consider religion?

○ Very important

○ Somewhat important

○ Not too important

○ Not at all important

Appendix E Software statement

The entire analysis was run under OS X 10.15.3 using R version 3.6.2 (R Core Team, 2017). In the empirical analysis, we made use of the following R software packages: Arnold (2019); Auguie (2017); Bache and Wickham (2014); Bates et al. (2015); Bates and Maechler (2019); Berger et al. (2017); Blair and Imai (2010); Comtois (2020); Dahl et al. (2019); Eady (2017); Firke (2020); Gelman and Su (2018); Genz and Bretz (2009); Genz et al. (2020); Grolemund and Wickham (2011); Henry and Wickham (2019, 2020); Hlavac (2018); Müller and Wickham (2019); Neuwirth (2014); Ooms (2019); Robinson and Hayes (2019); Rudis (2019); Venables and Ripley (2002); Wickham (2016); Wickham and Miller (2018); Wickham (2019*a*); Wickham and Miller (2019); Wickham and Bryan (2019); Wickham (2019*b*); Wickham and Henry (2019); Wickham et al. (2019, 2020); Zeileis (2004, 2006)

References

Arnold, Jeffrey B. 2019. *ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'.* R package version 4.2.0.
URL: *https://CRAN.R-project.org/package=ggthemes*

Auguie, Baptiste. 2017. *gridExtra: Miscellaneous Functions for "Grid" Graphics.* R package version 2.3.
URL: *https://CRAN.R-project.org/package=gridExtra*

Bache, Stefan Milton and Hadley Wickham. 2014. *magrittr: A Forward-Pipe Operator for R.* R package version 1.5.
URL: *https://CRAN.R-project.org/package=magrittr*

Bates, Douglas, Martin Mächler, Ben Bolker and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67(1):1–48.

Bates, Douglas and Martin Maechler. 2019. *Matrix: Sparse and Dense Matrix Classes and Methods.* R package version 1.2-18.
URL: *https://CRAN.R-project.org/package=Matrix*

Berger, Susanne, Nathaniel Graham and Achim Zeileis. 2017. Various Versatile Variances: An Object-Oriented Implementation of Clustered Covariances in R. Working Paper 2017-12 Working Papers in Economics and Statistics, Research Platform Empirical and Experimental Economics, Universität Innsbruck.
URL: *http://EconPapers.RePEc.org/RePEc:inn:wpaper:2017-12*

Blair, Graeme and Kosuke Imai. 2010. "list: Statistical Methods for the Item Count Technique and List Experiment." Available at The Comprehensive R Archive Network (CRAN).
URL: *https://CRAN.R-project.org/package=list*

Blair, Graeme and Kosuke Imai. 2012. "Statistical Analysis of List Experiments." *Political Analysis* 20(1):47–77.

Comtois, Dominic. 2020. *summarytools: Tools to Quickly and Neatly Summarize Data.* R package version 0.9.5.
URL: *https://CRAN.R-project.org/package=summarytools*

Dahl, David B., David Scott, Charles Roosen, Arni Magnusson and Jonathan Swinton. 2019. *xtable: Export Tables to LaTeX or HTML.* R package version 1.8-4.
URL: *https://CRAN.R-project.org/package=xtable*

Eady, Gregory. 2017. *misreport: Statistical Analysis of Misreporting on Sensitive Survey Questions.* R package version 0.1.1.
URL: *https://CRAN.R-project.org/package=misreport*

Firke, Sam. 2020. *janitor: Simple Tools for Examining and Cleaning Dirty Data.* R package version 1.2.1.
URL: *https://CRAN.R-project.org/package=janitor*

Gelman, Andrew and Yu-Sung Su. 2018. *arm: Data Analysis Using Regression and Multilevel/Hierarchical Models.* R package version 1.10-1.
URL: *https://CRAN.R-project.org/package=arm*

Genz, Alan and Frank Bretz. 2009. *Computation of Multivariate Normal and t Probabilities.* Lecture Notes in Statistics Heidelberg: Springer-Verlag.

Genz, Alan, Frank Bretz, Tetsuhisa Miwa, Xuefei Mi, Friedrich Leisch, Fabian Scheipl and Torsten Hothorn. 2020. *mvtnorm: Multivariate Normal and t Distributions.* R package version 1.0-12.
URL: *https://CRAN.R-project.org/package=mvtnorm*

Grolemund, Garrett and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40(3):1–25.
URL: *http://www.jstatsoft.org/v40/i03/*

Henry, Lionel and Hadley Wickham. 2019. *purrr: Functional Programming Tools*. R package version 0.3.3.
URL: *https://CRAN.R-project.org/package=purrr*

Henry, Lionel and Hadley Wickham. 2020. *rlang: Functions for Base Types and Core R and 'Tidyverse' Features*. R package version 0.4.4.
URL: *https://CRAN.R-project.org/package=rlang*

Hlavac, Marek. 2018. *stargazer: Well-Formatted Regression and Summary Statistics Tables*. Bratislava, Slovakia: Central European Labour Studies Institute (CELSI). R package version 5.2.2.
URL: *https://CRAN.R-project.org/package=stargazer*

Müller, Kirill and Hadley Wickham. 2019. *tibble: Simple Data Frames*. R package version 2.1.3.
URL: *https://CRAN.R-project.org/package=tibble*

Neuwirth, Erich. 2014. *RColorBrewer: ColorBrewer Palettes*. R package version 1.1-2.
URL: *https://CRAN.R-project.org/package=RColorBrewer*

Ooms, Jeroen. 2019. *writexl: Export Data Frames to Excel 'xlsx' Format*. R package version 1.2.
URL: *https://CRAN.R-project.org/package=writexl*

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rivers, Douglas. 2006. "Sample matching: Representative sampling from internet panels." *Polimetrix White Paper Series* .

Robinson, David and Alex Hayes. 2019. *broom: Convert Statistical Analysis Objects into Tidy Tibbles*. R package version 0.5.3.
URL: *https://CRAN.R-project.org/package=broom*

Rudis, Bob. 2019. *hrbrthemes: Additional Themes, Theme Components and Utilities for 'ggplot2'*. R package version 0.6.0.
URL: *https://CRAN.R-project.org/package=hrbrthemes*

Stier, Sebastian, Johannes Breuer, Pascal Siegers and Kjerstin Thorson. 2019. "Integrating Survey Data and Digital Trace Data: Key Issues in Developing an Emerging Field." *Social Science Computer Review* p. 0894439319843669.

Venables, W. N. and B. D. Ripley. 2002. *Modern Applied Statistics with S*. Fourth ed. New York: Springer. ISBN 0-387-95457-0.
URL: *http://www.stats.ox.ac.uk/pub/MASS4*

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.
URL: *https://ggplot2.tidyverse.org*

Wickham, Hadley. 2019*a*. *forcats: Tools for Working with Categorical Variables (Factors)*. R package version 0.4.0.
URL: *https://CRAN.R-project.org/package=forcats*

Wickham, Hadley. 2019*b*. *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.4.0.
URL: *https://CRAN.R-project.org/package=stringr*

Wickham, Hadley and Evan Miller. 2018. *haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files*. R package version 1.1.1.

Wickham, Hadley and Evan Miller. 2019. *haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files*. R package version 2.2.0.
URL: *https://CRAN.R-project.org/package=haven*

Wickham, Hadley and Jennifer Bryan. 2019. *readxl: Read Excel Files*. R package version 1.3.1.
URL: *https://CRAN.R-project.org/package=readxl*

Wickham, Hadley and Lionel Henry. 2019. *tidyr: Tidy Messy Data*. R package version 1.0.0.
URL: *https://CRAN.R-project.org/package=tidyr*

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo and Hiroaki Yutani. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4(43):1686.

Wickham, Hadley, Romain François, Lionel Henry and Kirill Müller. 2020. *dplyr: A Grammar of Data Manipulation*. R package version 0.8.4.
URL: *https://CRAN.R-project.org/package=dplyr*

Zeileis, Achim. 2004. "Econometric Computing with HC and HAC Covariance Matrix Estimators." *Journal of Statistical Software* 11(10):1–17.

Zeileis, Achim. 2006. "Object-Oriented Computation of Sandwich Estimators." *Journal of Statistical Software* 16(9):1–16.