Research note — Version: June 13, 2020

Citizens' Preferences for Online Hate Speech Regulation in the United States and Germany

Simon Munzert[1], Richard Traunmüller[2], Pablo Barberá[3], Andrew Guess[4], and JungHwan Yang[5]

[1]Assistant Professor of Data Science and Public Policy, Hertie School
[2]Professor of Empirical Democracy Research, University of Mannheim
[3]Associate Professor of Political Science and International Relations, University of Southern California
[4]Assistant Professor of Politics and Public Affairs, Princeton University
[5]Assistant Professor of Communication at the University of Illinois at Urbana-Champaign

Abstract. To openly express one's views is a fundamental right in any liberal democracy. However, in the age of social media, the questions of what is allowed to say and how public discourse should be regulated are ever more contested. We present a pre-registered study to analyze citizens? preferences for online hate speech regulation. We construct vignettes in forms of social media posts, mimicking actual cases of hate speech to bolster external validity, that vary along key dimensions of hate speech regulation. Respondents are asked to judge the posts with regards to perceived offensiveness and hatefulness as well as to actions that should be taken by the platform providers and other consequences the sender of hate speech should face. The experiment is embedded in nationally representative online panels in the US and Germany, which allows us to analyze context- as well as individual-level determinants for online hate speech regulation. The results indicate both convergence and divergence in how people in different contexts think about hate speech regulation, which has implications for its public legitimacy.

Introduction

To openly express one's views is one of the most fundamental rights in any liberal democracy and anchored in Article 19 in the Universal Declaration of Human Rights. But even this right has its limits when acts of hate speech impinge upon the dignity and rights of others. As a result of this tension, the questions of what is allowed to say and how public discourse should be regulated are highly contested. Underlying these debates are world views regarding the potential harm of speech and its suppression (Barendt, 2009; Massaro, 1990).

A global survey has recently found strong variation in the extent citizens in 64 countries support free expression (Wike and Simmons, 2015). However, while policymakers around the world have started to take action on hate speech regulation and online social media companies struggle to develop and implement community standards on hate speech, little is known about what exactly the users of these platforms—and the public in general—deem hateful speech. As importantly, citizens' preferences regarding action in response to hateful content by platform providers, governments, or the civil society are not well understood yet. Any form of action is further complicated by wildly varying norms and cultures of free speech across countries and regions.

Motivated by these developments, our research aims to study what citizens around the world think about the limits of public speech and how they balance the goals of freedom, group equality, or the prevention of harm. In particular, we are guided by the following overarching questions: 1. What do citizens deem acceptable or unacceptable speech in online public discourse?, 2. How do citizens want online speech to be handled and regulated?, and 3. Which individual factors determine these preferences, and how does cultural context matter?

What shapes citizens' preferences for online hate speech regulation?

We believe that sensitivity for controversial content as well as preferences towards hate speech regulation are influenced by (a) content- and context-specific factors of the content itself, (b) individual characteristics, and (c) contextual factors such as existing legislation, planned policies, and cultural norms of free speech.

With regards to content-specific factors, we hypothesize that controversial content is considered relatively more hateful and offensive when the issue involves treatment of minorities, such as Muslim immigrants, and relatively less hateful and offensive when it involves treatment of broad groups, such as political movements. Furthermore, we expect severity to matter: violent content should be regarded more problematic that insults, vilification, or discrimination. In addition, we also study the influence of target and sender identities, the message context, the addressing scope, and target reactions.

Moreover, we believe that hate speech sensitivity and preferences for regulation are conditional on respondent characteristics. In particular, we hypothesize that citizens will be biased in favor of their own group identity defined by characteristics such as gender, ethnicity, political ideology, or religion Grant and Rudolph (2003). In addition, we test for the baseline impact of factors that have been shown to structure attitudes towards free speech and its regulation in general. These factors include political ideology (Lalonde et al., 2000; Gross and Kinder, 1998; Chong, 2006; Suedfeld et al., 1994), gender (Downs and Cowan, 2012), age (Lambe, 2004), ethnicity (Gross and Kinder, 1998; Chong, 2006), and cultural context.

Design

To study individual perceptions and preferences, we use vignettes that are constructed in a way that mimics posts on a popular social media platform (here: Facebook). Irrelevant features of the message, such as time stamp or features to interact with it, are dropped. Only features that represent relevant attributes of the vignettes are kept. These attributes cover issues, sender as well as target characteristics, and sender message's and target message's characteristics. Table 1 provides an overview of the attributes and attribute levels.

We use a set of questions that respondents answer for each vignette. They provide measures of (a) how offensive or hateful a respondent sees the respective post and (b) what consequences the posts or the author of the post should face. Actions that are offered include measures the platform provider should take (no action, delete post, prohibit sender to post on target's timeline, temporarily ban sender, permanently ban sender) and others that the sender's employer or the law enforcement should take (no further action, lose job, fine, jail). To estimate the causal effects, we calculate *average marginal component effects* (AMCE) using linear regression with clustered standard errors (Hainmueller et al., 2013). The experiment, in which each respondent was shown eight of these vignettes, was embedded in two panel surveys fielded on about 1,300 respondents each recruited for the YouGov U.S. and German Pulse panel, launched in early 2019. The analysis was pre-registered at EGAP (https://egap.org/registration/5944).

Results

Figure 1 provides a summary of the core evaluation scales for the social media posts. The perceived offensiveness and hatefulness is strongly related (Pearson correlation of $r = .79$ in the German sample and $r = .89$ in the US sample). A majority of the evaluations consider the shown social media posts to be at least "somewhat offensive" or "somewhat hateful" by the respondents. While the distributions are tilted towards evaluations rating the posts as

Table 1. Vignette attributes and levels

Attribute	**Levels**
Issue ∈	Muslim immigrants, Women, Ideological Left, Ideological Right
Target's identity ∈	male/female, Muslim/non-Muslim, liberal anonymous/conservative anonymous
Target's initial message ∈	identification with target group, support of target group
Sender's identity ∈	male/female, Muslim/non-Muslim, liberal anonymous/conservative anonymous
Sender message's target ∈	"You...", "All...", "Most...", "Extreme..."
Sender's message ∈	violence (moderate/extreme), insult (moderate/extreme), discrimination (moderate/extreme), vilification (moderate/extreme)
Target's reaction ∈	appealing to norms, counter-aggression, platform action (blocking, reporting), none

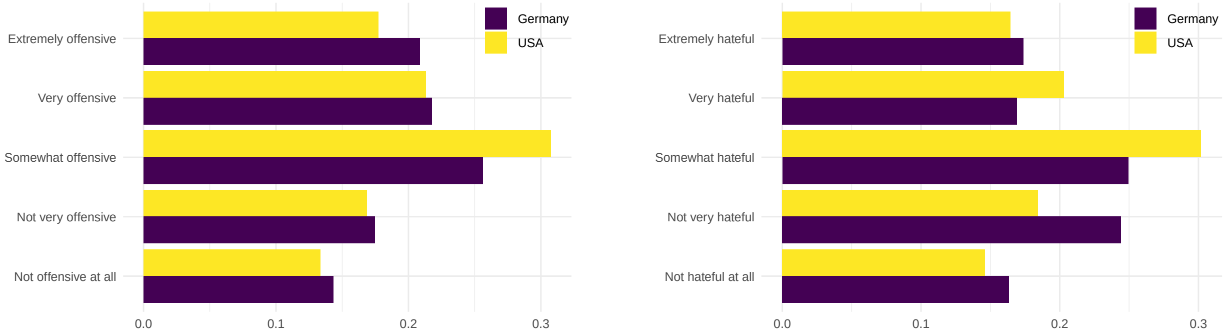

Figure 1. Respondents' perceptions of offensiveness and hatefulness of shown social media posts (shares of all posts shown).

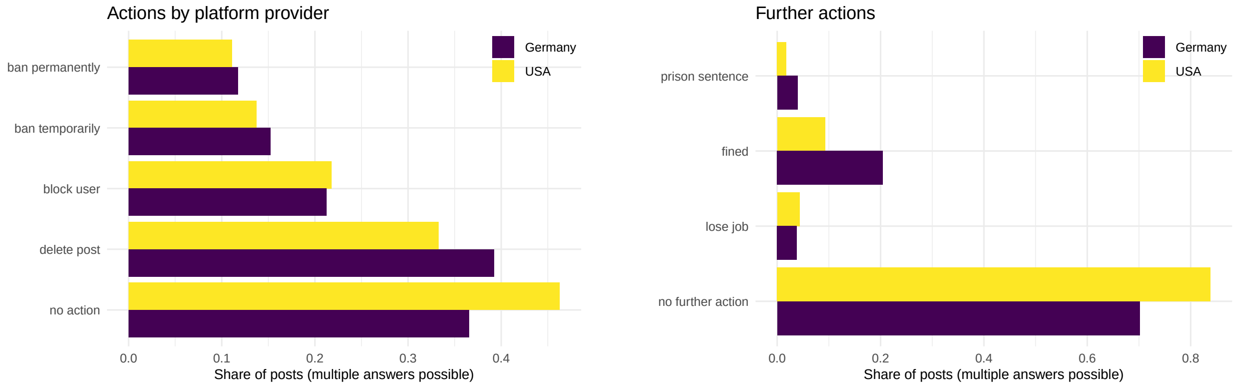

Figure 2. Respondents' preferred actions by platform provider and additional actions in response to of shown social media posts (shares of all posts shown).

offensive or hateful (about 1/5 of the posts are even considered "extremely offensive/hateful"), there is substantial variation in the ratings. These findings indicate that the posts, which were designed to be on the spectrum between potentially controversial to strongly offensive/hateful, cover the entire evaluation scales.

Figure 2 reports the share of posts for which a particular action by the platform provider or further consequences were chosen as appropriate. For roughly 40% (US sample: 45%) of the posts, the respondents saw no need for action by the platform provider. On the other side of the spectrum, for 12% of the posts the respondents would have liked to see the sender of the message to be banned permanently from the platform. The other options are somewhere in between, which the order mirrors a plausible ladder of escalation. With regard to further, possibly legal consequences, about 70% (US: 82%) of all social media posts where deemed as not requiring any further action. However, in about 5% of the cases respondents thought the sender should lose their job, and for 20% (US: 10%) they suggested a fine. A prison sentence was almost never considered an appropriate sanction (2%). While the country-level differences provide suggestive evidence that cultural context matters for perceptions and preferences towards hateful content, these differences are less pronounced in the following analyses, which is why we present pooled results.

Finally, we report the AMCEs of content and context features of the created messages (Figure 3) as well as selected respondent characteristics (Figure 4) on preferred actions. The effects on perceptions of offensiveness and hatefulness follow very similar patterns (not shown). Three consistent patterns emerge: First, the type and severity of hate speech matters a lot; the violent and insulting posts as well as the more extreme forms of all categories trigger higher average preferences for action. Second, hate speech that is directed towards women and Muslims is consistently perceived as more problematic than content directed towards partisan camps. Third, all other factors have remarkably little influence on peoples' perceptions and preferences. In the analysis of respondent characteristics, two robust patterns emerge: Women as well as people who self-identify as left-leaning or left are on average consistently more likely to support action against the controversial posts.

Discussion and conclusion

We conclude with a brief summary and discussion of the key findings. **Type and degree of hate speech matters.** People perceive more extreme violence and insults as more offensive and hateful and are willing to take action. This also implies that people are able to make nuanced judgement calls and seem to apply heuristics that are consistent with theoretical accounts of hate speech types and their severity. **Content matters.** Some topics (hate speech against women, Muslims) are perceived as more problematic than others (hate speech against partisans). This is interesting insofar as the actual message content was kept mostly constant over the different topics. Put bluntly, a slur against women is considered more problematic than a slur against Republicans. **Message context less important.** We find sender or target attributes to exert little effect on the outcomes. **Preferences vary systematically across citizen characteristics.** Furthermore, in analyses not shown here we find evidence for substantive bias in favor of people's own group identity.

We suggested an efficient way to elicit citizens' preferences on hate speech regulation using concrete examples not abstract considerations. Neither recent and ongoing government nor platform efforts to combat hateful and offensive content on the internet seems to be informed by public preferences in any meaningful way. We consider learning about public preferences relevant for designing regulations that are met by wide acceptance from general publics. We hope that the results and methods we presented provide a fruitful starting point for a more evidence-based approach to online content regulation.

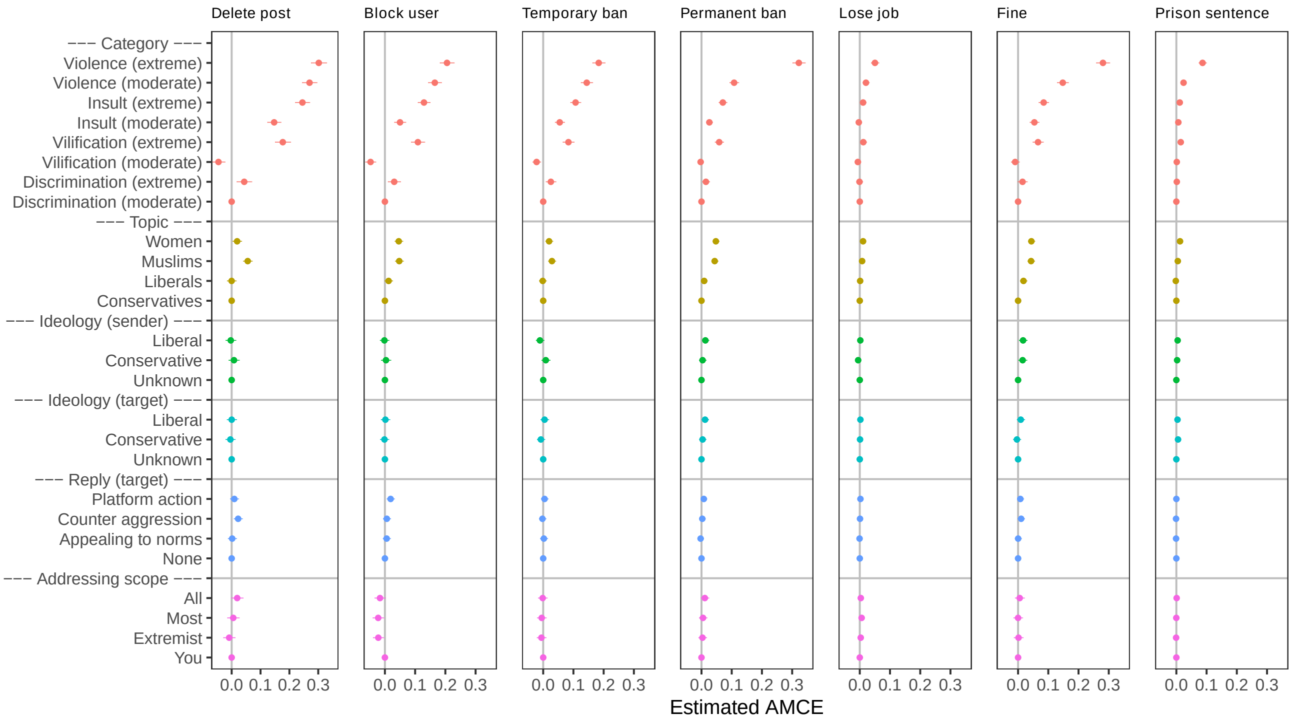

Figure 3. Estimated average marginal component effects of content and context characteristics of social media hate speech vignettes on citizens' preferred platform-side and other action.

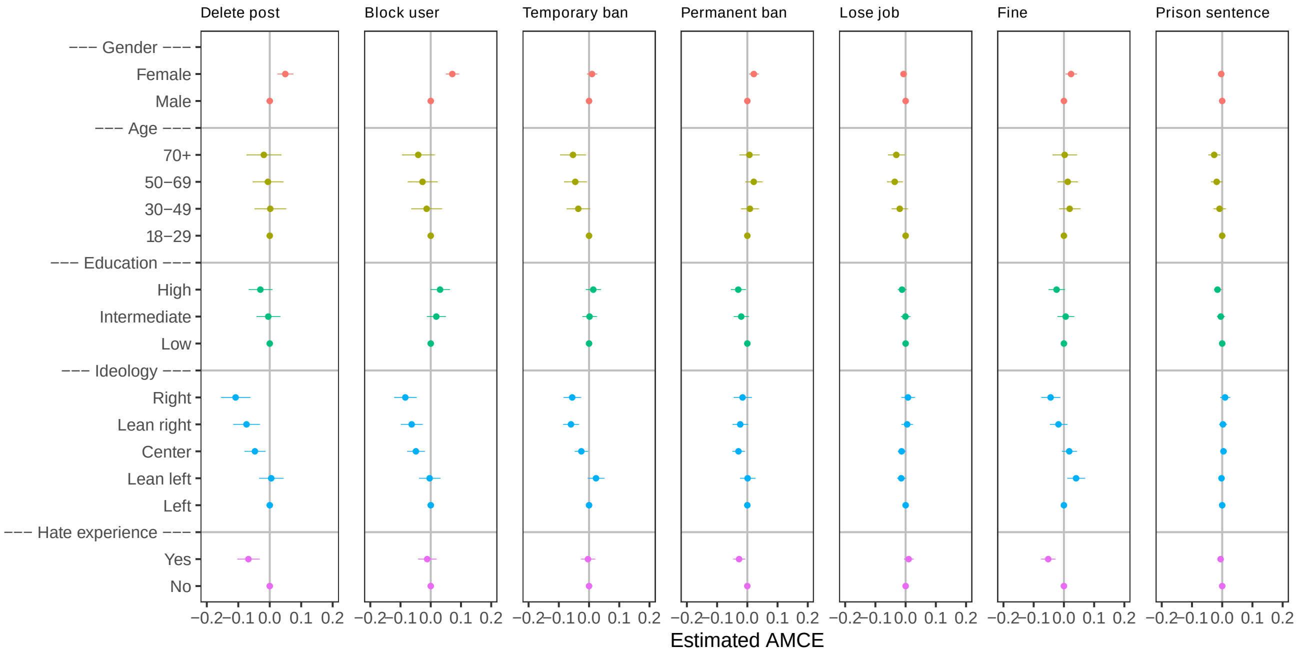

Figure 4. Estimated average marginal component effects of respondent characteristics on citizens' preferred platform-side and other action.

References

Barendt, E. (2009): "Balancing freedom of expression and privacy: the jurisprudence of the Strasbourg Court," *Journal of Media Law*, 1, 49–72.

Chong, D. (2006): "Free speech and multiculturalism in and out of the academy," *Political Psychology*, 27, 29–54.

Downs, D. M. and G. Cowan (2012): "Predicting the importance of freedom of speech and the perceived harm of hate speech," *Journal of applied social psychology*, 42, 1353–1375.

Grant, J. T. and T. J. Rudolph (2003): "Value conflict, group affect, and the issue of campaign finance," *American Journal of Political Science*, 47, 453–469.

Gross, K. A. and D. R. Kinder (1998): "A collision of principles? Free expression, racial equality and the prohibition of racist speech," *British Journal of Political Science*, 28, 445–471.

Hainmueller, J., D. J. Hopkins, and T. Yamamoto (2013): "Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments," *Political Analysis*, 22, 1–30.

Lalonde, R. N., L. Doan, and L. A. Patterson (2000): "Political correctness beliefs, threatened identities, and social attitudes," *Group Processes & Intergroup Relations*, 3, 317–336.

Lambe, J. L. (2004): "Who wants to censor pornography and hate speech?" *Mass Communication & Society*, 7, 279–299.

Massaro, T. M. (1990): "Equality and freedom of expression: The hate speech dilemma," *Wm. & Mary L. Rev.*, 32, 211.

Suedfeld, P., G. D. Steel, and P. W. Schmidt (1994): "Political Ideology and Attitudes Toward Censorship 1," *Journal of Applied Social Psychology*, 24, 765–781.

Wike, R. and K. Simmons (2015): "Global Support for Principle of Free Expression, but Opposition to Some Forms of Speech," *Pew Research Center Report*, November 18.