Mueller, Karsten; Schwarz, Carlo

Article

Online hate speech, offline harm, and the case for content moderation

EconPol Forum

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Mueller, Karsten; Schwarz, Carlo (2025) : Online hate speech, offline harm, and the case for content moderation, EconPol Forum, ISSN 2752-1184, CESifo GmbH, Munich, Vol. 26, Iss. 4, pp. 41-46

This Version is available at:
https://hdl.handle.net/10419/334429

Policy Debate of the Hour

Online Hate Speech, Offline Harm, and the Case for Content Moderation

Key Messages

- Hateful online activity not only reflects underlying tensions but affects real-world behavior
- Moderation policies should be based on existing laws to avoid new speech restrictions and dismiss claims of digital censorship
- Germany's NetzDG regulation had instant online and offline effects, disrupting the channels that turn hatred into violence
- Removing posts may distort online speech, but a "softening" of hateful expression through generative language models could reduce toxicity while minimizing distortions
- Moderation rules must be clear and narrow to limit compliance costs without entrenching incumbents

Karsten Müller is Presidential Young Professor and Assistant Professor of Finance at the National University of Singapore's Business School. He is also a Research Fellow at the NUS Risk Management Institute.

Carlo Schwarz is an Assistant Professor in the Department of Economics at Bocconi University. His research combines causal inference strategies with techniques from text analysis, machine learning, and data science.

Policymakers around the world are grappling with how to curb harmful online speech without unduly infringing on free expression. The core motivation for content moderation is straightforward: while freedom of speech is a bedrock principle of open societies, it is not an absolute right – it ends where the exercise of that freedom causes harm to others.

In the digital age, a key concern is that unchecked hate speech and incitement on social media may translate into real-world violence and strain on societal cohesion. This article explores the empirical evidence linking online hate to offline harm. We also discuss the case for online

content moderation and what goals it should serve. The central argument is that targeted moderation of extreme content can reduce violence and make online spaces safer without undermining core principles governing freedom of expression.

Free speech is essential in a democracy, but it has well-recognized limits when it conflicts with the safety and rights of others (e.g., Mill 1859). The classic example is that one cannot shout "Fire!" in a crowded theatre, because it is likely to incite panic and thereby cause harm to others. In the same vein, almost all legal systems at least partly forbid threats of violence, harassment, defamation, or incitement to imminent lawless action, recognizing that such speech has the potential to inflict tangible harm.

The advent of social media has significantly lowered the entry barriers for individuals to share their views online, and some have followerships that far outstrip the viewership of major TV shows. The issue is that, when abuse, extremism, and falsehoods are widespread on online platforms, it drives away more civil users and thereby distorts the public debate. It also undermines the very free exchange of ideas that free speech is meant to facilitate by creating an environment in which civil debate is all but impossible.

From Online Hate to Offline Violence

Extreme forms of online hate speech undermine the online debate, as hate speech often targets people based on race, religion, gender, or other group characteristics, and can create an atmosphere of fear. A critical question for policymakers is whether online hate speech translates into real-world violence offline – and a growing body of empirical research indicates that it does. In recent years, economists and social scientists have leveraged data from social media platforms and crime statistics to quantify this link. Two studies by Müller and Schwarz (2021, 2023) provide evidence from Germany and the United States.

Facebook and Anti-refugee Incidents in Germany

In Müller and Schwarz (2021), we examined the relationship between anti-refugee hate speech on Facebook in Germany and hate crimes against refugees. Spikes in anti-refugee posts on Facebook systematically coincide with increases in violent crimes against refugee populations, particularly in areas with high Facebook usage. To establish a causal connection between anti-refugee incidents and online hate speech, we exploit both local internet and Germany-wide Facebook outages, which cut off individuals from exposure to online anti-refugee content without affecting other factors.

Figure 1 summarizes one of the main findings from the paper. The blue dots in panel A show that, in towns where the ratio of Facebook users who follow the page of the right-wing party Alternative für Deutschland (AfD) relative to the overall population is equal to or above the median, Facebook posts about refugees are highly correlated with the probability of an offline attack on refugees. In panel B, the corresponding blue dots show a somewhat less steep correlation for towns with a below-median ratio.

The key finding can be seen when comparing the blue dots to the orange diamonds. During internet outages, the correlation between online hate and offline violence disappears. In other words, when the potential perpetrators of hate crimes are no longer able to access social media to the same extent, the link between hateful rhetoric on Facebook and real-life hate crimes vanishes. The most plausible interpretation of this pattern is that hateful activity on Facebook can indeed cause hate crimes, and does not only reflect underlying tensions. This evidence was the first to demonstrate a concrete causal link between social media content and real-world acts of violence.

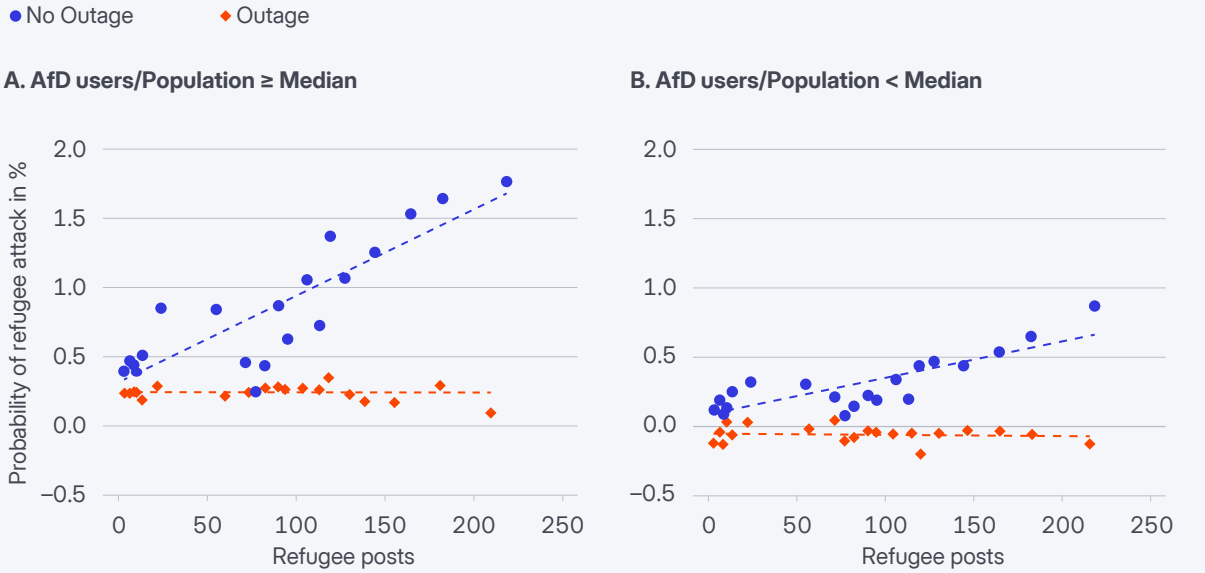
Twitter and Anti-Hispanic/Muslim Hate Crime in the United States

In a follow-up study, titled "From Hashtag to Hate Crime," we turned to the US, focusing on Twitter (now "X") and anti-minority sentiment. In particular, we asked whether surges of hateful rhetoric on Twitter lead to more hate-fueled incidents offline. In our empirical analysis, we used variation in the adoption of Twitter across US counties, which was propelled partly by the South by Southwest (SXSW) tech conference that gave Twitter an initial boost in certain counties. We then analyzed the connection between Twitter usage and hate crimes against Muslims and other minorities, particularly following inflammatory tweets by Donald Trump during his first presidential campaign and presidency.

The data suggests a clear pattern. Counties with higher Twitter penetration experienced significantly larger increases in hate crimes targeting minorities in the aftermath of Donald Trump's presidential run, particularly at times when Donald Trump tweeted anti-minority content. For example, following major terrorist attacks that spurred anti-Muslim hashtags and vitriol on Twitter, areas where Twitter usage was higher saw a disproportionate rise in anti-Muslim hate crimes. The evidence suggests a propagation effect: online hate speech contributes to prejudice and encourages a small subset of already radicalized individuals to commit violent acts.

Figure 1

Anti-Refugee Posts and Anti-Refugee Incidents



Note: Panel A shows a binned scatter plot of local posts on the AfD Facebook page as a function of the reports on internet outages in a given week. Panel B plots the average number of antirefugee attacks against our measure of antirefugee sentiment for municipalities above and below the median of AfD users/Population. The data in total contains 479,964 municipality-week observations (4,324 municipalities over 111 weeks). Refugee attacks are binned by 20 quantiles of refugee posts. We additionally split towns by whether they experience an internet outage in a given week (orange squares). The number of anti-refugee attacks is residualized with respect to population; hence, the number of attacks can be slightly below 0 in some bins.
Source: Müller and Schwarz (2021).

These findings are also consistent with other research linking Trump's rhetoric to changes in social norms. For instance, Bursztyn, Egorov, and Fiorin (2020) show that the election of Trump made individuals more willing to express xenophobic views. In other research, Bursztyn, Egorov, Enikolopov and Petrova (2024) provide evidence of a connection between social media and hate crime in Russia.

In sum, the empirical literature increasingly supports the idea that extreme online hate speech is not just harmless venting. Instead, it has the potential to incite or amplify real-world harm. This strengthens the case for some kind of content moderation, but it also raises the crucial question of whether such moderation efforts are effective in reducing the prevalence of hateful online content and its offline consequences.

Does Content Moderation Work?

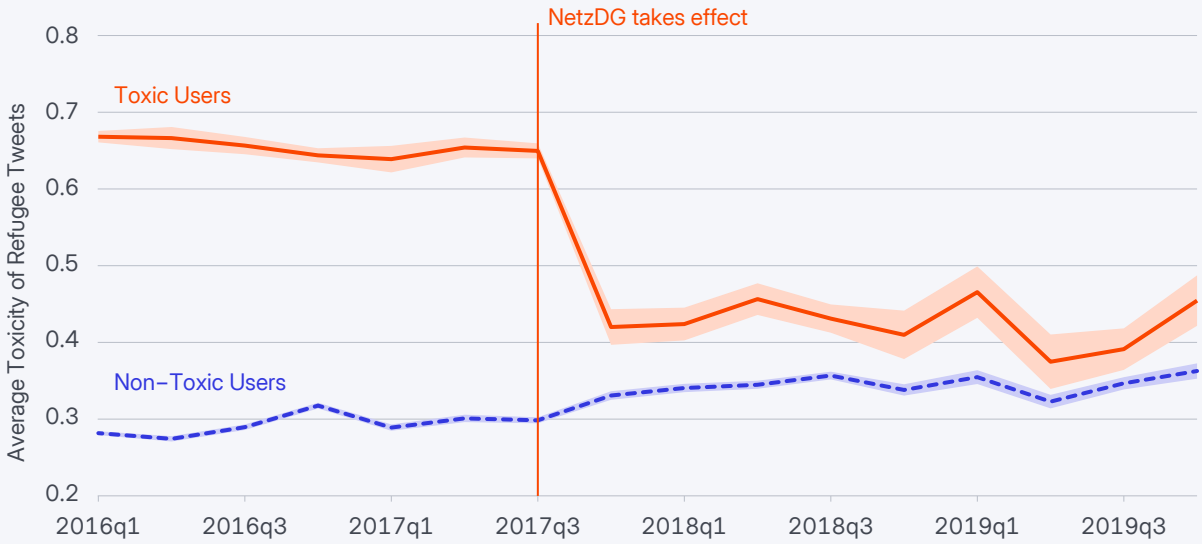
The crux of content moderation is that it should balance the ability of citizens to freely express their opinions with the need to uphold existing laws or rules preventing the incitement of violence (and perhaps more broadly, a certain level of civility in public debate). Social media users, in fact, broadly agree on taking down explicit incitements to violence and egregious hate speech attacking certain groups (e.g., Solomon, Hall, Hemmen and Druckman, 2024). As such, there is a popular mandate for platforms to eliminate content that calls for violent acts or contains dehumanizing hate speech and only few would insist that such content is a healthy component of public discourse.

Some of the first evidence of the effects of content moderation stems from Jiménez (2023). His study shows that, when hateful posts are randomly removed from Twitter, the users who are the targets of these posts increase their engagement, but those publishing them do not lay off their hateful behavior. The implication is that content moderation does not, in fact, deter users, but rather increases advertising revenue by making platforms more engaging for people who would otherwise be verbally attacked.

In a recent paper – Jiménez, Müller and Schwarz (2025) – we turn our attention to the offline impacts of content moderation by studying the effects of Germany's Netzwerkdurchsetzungsgesetz (Network Enforcement Act or NetzDG for short). Passed in 2017, the law's explicit aim was to

Figure 2

The Effect of the NetzDG on Online Toxicity



Note: This figure plots the average toxicity of tweets containing the word refugee ("Flüchtling"). We split Twitter users based on the toxicity of their refugee tweets in the pre-period. "Toxic Users" are those whose tweets before the NetzDG law came into effect were on average above the 90th percentile of the toxicity distribution. "Non-Toxic Users" are all remaining users. The shaded areas indicate 95 percent confidence intervals for the means.
Source: Jiménez et al. (2025). © ifo Institute

fight hate crime, criminally punishable fake news, and other unlawful content on social networks more effectively by obligating platforms to remove such content quickly. More specifically, NetzDG required large social media platforms to promptly remove content that violates German law (such as criminal hate speech), and non-compliance could incur fines up to EUR 50 million. Crucially, the NetzDG did not create new speech restrictions; it simply made clear that the German Criminal Code provisions on hate speech, public incitement, and insults apply online, and put the onus on platforms to enforce those existing laws.

Germany's NetzDG law, therefore, provides a rare real-world test of content moderation's effects. In our study, we found that NetzDG had an immediate, visible effect on Germany's online discourse. Figure 2 shows the average tweet toxicity of users conditional on their pre-period level of toxicity. The figure shows that hate-laden posts – especially refugee-related tweets from far-right users – fell significantly once the law took effect. This echoes findings by Andres and Slivko (2021) that German extremists became noticeably less abusive than their Austrian counterparts after 2018.

Crucially, this effect did not drive people off the major platforms. Overall posting volumes and topic concentration barely moved, and use of Facebook and Twitter even ticked up after the law, suggesting that stricter rules made mainstream users more rather than less willing to engage while leaving heavy "toxic" posters active but more civil.

Additionally, we show that NetzDG also had offline consequences. In municipalities with the greatest pre-law exposure to far-right social media content, anti-refugee hate crimes dropped about one percent relative to trend; a country-level synthetic-control analysis shows a parallel national decline in hate crime overall. We attribute this decline to a disruption of extremists' ability to mobilize online. We also provide evidence that NetzDG did not suddenly soften personal prejudices, but by deleting rallying calls and propaganda, it broke the coordination channels that turn online hatred into offline violence.

In a sense, the German experience offers a proof of concept: enforcing existing legal standards online, backed by penalties, can lead to less toxic speech and fewer hate crimes. Content moderation policy can thus achieve societal benefits, although it should be clear that it cannot address deep-seated sources of hatred. For policymakers, NetzDG underscores that moderation need not mean stifling political debate, which we find continued, just with a reduction in overt hate and violent incitement. In another

paper (Müller and Schwarz 2024), we provide evidence of a reduction in the toxicity of Trump's followers after the removal of his Twitter account.

Understanding the Costs of Content Moderation

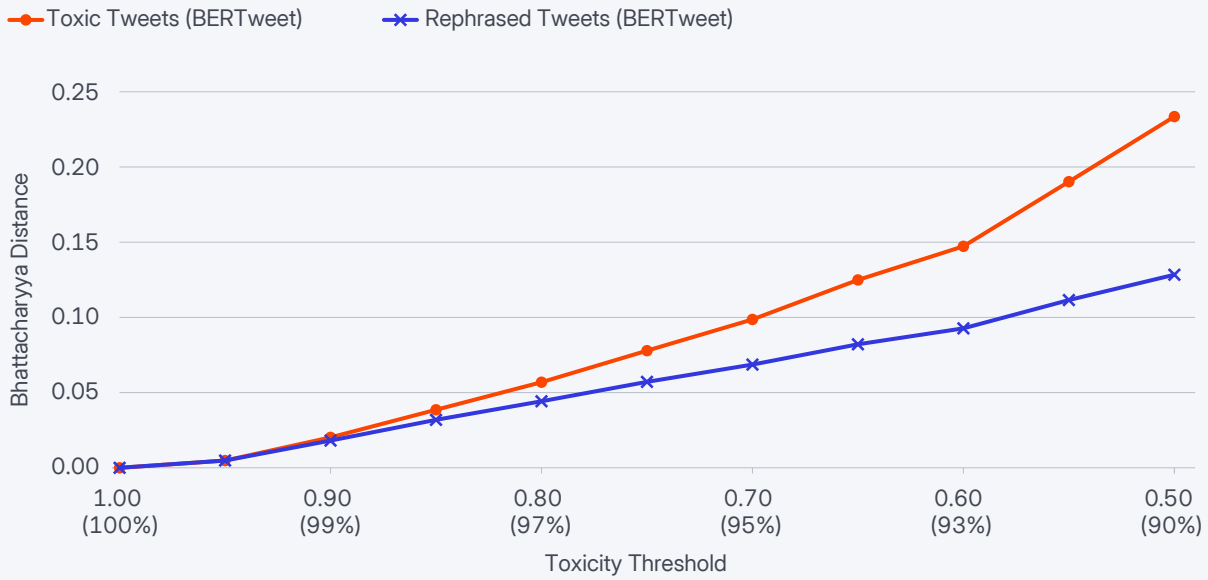
Even if we accept the positive effects of content moderation based on the above studies, policymakers face a clear dilemma. If removing toxic posts improves the online conversation, it can also inadvertently distort the remaining content and narratives in ways that might be undesirable, especially if potentially important information or viewpoints are removed.

This challenge is explored in recent research by Habibi, Hovy and Schwarz (2025), who develop a methodology for quantifying to what extent deleting toxic content shifts the overall content of online discussions. Based on a large dataset of 5 million political tweets, they compare the distribution of content with and without the toxic tweets. The key finding is that removing toxic speech does indeed distort online discourse, not only mechanically by deleting certain hate-filled words, but rather because toxic posts often contain some genuine topical content (e.g., political opinions, albeit couched in slurs or profanities). Removing them entirely also removes the viewpoints or information they carried (even if delivered offensively), thereby shifting the topical balance of the remaining content. This evidence confirms precisely the fears of those opposing content moderation: higher perceived civility by some may come with a censorship of discourse.

Habibi et al. (2025) show that such fears can be at least reduced using potential alternative moderation strategies. In particular, we experiment with using generative language models to rephrase toxic content into less toxic terms rather than simply deleting it. In other words, an automatic filter would take a hateful post and generate a cleaned-up version that preserves the substantive point but removes the slurs and violent rhetoric, thus leaving the topic of conversation in place. The study demonstrates that this approach, essentially a "softening" of the delivery of toxic content, can reduce toxicity scores while minimizing distortions in the overall content distribution. This result is visualized in Figure 3, which shows the evolution of our measure of content distortions (y-axis) as a function of the extent of content moderation. The orange line represents the distortions from the removal of tweets, while the blue line shows the distortions if toxic tweets are instead replaced with a rephrased version.

Figure 3

Content Moderation and Distortions to Online Content



Note: The figure shows changes in the embedding space, created based on the BERTweet model, for two different content moderation strategies. The sample consists of 5 million US Twitter users. The orange line shows the Bhattacharyya distance if toxic tweets are removed from the sample. The blue line shows the Bhattacharyya distance if toxic tweets are rephrased. The x-axis indicates the toxicity threshold above which tweets are removed or rephrased. The figure also reports the share of tweets that remain in the sample in parentheses below.
Source: Habibi et al. (2025).

Policy Implications: Keep Moderation Feasible, Simple, and Transparent

Taking stock, empirical research clearly provides a rationale for democratic societies to implement some kind of content moderation to stem the harmful offline effects of online hatred. Three findings in particular stand out. First, extreme hate speech is easier to spot and almost universally rejected. Second, a tiny pool of highly radical users generates most toxic material and violence, so that targeting them yields outsized gains for the majority. Third, rephrasing toxic posts might bring some public-safety benefits with likely negligible losses to free-speech values; given they are rare, even deleting some toxic posts may not seriously impair free debate. In our view, the case for content moderation is straightforward: when based on clear, transparent, and technologically aided principles, moderation can curb the most dangerous speech without stifling viewpoints in a democracy.

Sound moderation policies should begin with a simple rule: whatever is illegal to say in a newspaper or on a street corner should be just as impermissible online. Anchoring policies to existing laws on threats, hate incitement, child-abuse material, and defamation avoids inventing new speech restrictions and undercuts claims of digital "censorship." While several restrictions on speech in European countries should be reviewed (e.g., §188 of the German Criminal Code (StGB), which punishes insults against politicians), most legal codes provide reasonable restrictions consistent with the harm principle.

The main challenge is execution in the online space. Mass online platforms must process millions of posts a day, so moderation rules have to be clear and narrow. Possible rules could include "no direct threats" or "no ethnic slurs," but they should avoid vague bans on "offensive content." Such clarity lets algorithms flag the most extreme hateful content that has historically been associated with offline violence, and an algorithmic implementation would give users confidence that enforcement is rule-based rather than arbitrary. Most importantly, content moderation would signal that authorities are serious about upholding offline law online.

Keeping rules simple also limits compliance costs. Giants such as Meta can afford to hire thousands of reviewers; small platforms cannot. Tiered obligations, shared industry tools, or open-source filters can help newcomers meet legal duties without entrenching incumbents. Even then, mistakes are inevitable: some lawful posts will be removed and some unlawful ones missed. But at this point, the evidence is fairly unambiguous that the societal costs of allowing clearly hateful calls for violence to circulate online are not only against most people's wishes but also have offline consequences. By refusing to let a few toxic individuals poison the discourse, content moderation may thus make online spaces better marketplaces of ideas.•

References

Andres, R. and O. Slivko (2021), "Combating online hate speech: The impact of legislation on Twitter", *ZEW Discussion Papers*, (No. 21–103).

Bursztyn, L., G. Egorov, and S. Fiorin (2020), "From extreme to mainstream: The erosion of social norms", *American Economic Review* 110.11, 3522–3548.

Bursztyn, L., G. Egorov, R. Enikolopov, and M. Petrova (2024), "Social Media and Xenophobia: Theory and Evidence from Russia", *NBER Working Paper* 26567.

Habibi, M., D. Hovy, and C. Schwarz (2025), "The Content Moderator's Dilemma: Removal of Toxic Content and Distortions to Online Discourse", *arXiv Preprint* 2412.16114.

Jiménez-Durán, R. (2023), "The Economics of Content Moderation: Theory and Experimental Evidence from Hate Speech on Twitter", *SSRN Working Paper* 4590147.

Jiménez-Durán, R., K. Müller, and C. Schwarz (2025), "The Online and Offline Effects of Content Moderation: Evidence from Germany's NetzDG", *SSRN Working Paper* 4230296.

Mill, J. S. (1859), "On liberty", in *A selection of his works*, London: Macmillan Education UK, 1–147.

Müller, K. and C. Schwarz (2021), "Fanning the Flames of Hate: Social Media and Hate Crime", *Journal of the European Economic Association* 19(4), 2131–2167.

Müller, K. and C. Schwarz (2023), "From Hashtag to Hate Crime: Twitter and Antiminority Sentiment", *American Economic Journal: Applied Economics* 15(3), 270–312.

Müller, K. and C. Schwarz (2024), "The Effects of Online Content Moderation: Evidence from President Trump's Account Deletion", *SSRN Working Paper* 4296306.

Solomon, B. C., M. E. Hall, A. Hemmen, and J. N. Druckman (2024), "Illusory interparty disagreement: Partisans agree on what hate speech to censor but do not know it", *Proceedings of the National Academy of Sciences*, 121(39).