The Online and Offline Effects of Content Moderation: Evidence from Germany's NetzDG*

Rafael Jiménez Durán† Karsten Müller‡ Carlo Schwarz§

February 25, 2025

Abstract

Social media companies are under scrutiny for the prevalence of hateful content on their platforms, but there is little empirical evidence on the consequences of moderating such content. We study the online and offline effects of content moderation on social media using the introduction of Germany's "Network Enforcement Act" (NetzDG), which fines social media platforms for failing to remove hateful posts, as a natural experiment. We show that the NetzDG reshaped social media discourse: posts became less hateful, refugee-related content became less inflammatory, and the use of moderated platforms increased. Notably, the law did not significantly reduce the overall activity of toxic users or alter conversation topics. Offline, the NetzDG caused a 1% reduction in anti-refugee hate crimes for every standard deviation in far-right social media usage. Using a synthetic control approach, we document similar effects on overall hate crimes in Germany. In terms of mechanisms, we provide evidence that the NetzDG decreased hate crimes by reducing collective action rather than changing attitudes toward refugees.

Keywords: Social Media, NetzDG, Content Moderation, Hate Crime, Refugees, Germany
JEL Codes: L82, J15, O38.

*We are grateful to Leonardo Bursztyn, Ruben Durante, Fabrizio Germano, Ruben Enikolopov, Sophie Hatte, Ro'ee Levy, Sulin Sardoschau, Joshua Tucker, David Yang, Noam Yuchtman, Ekaterina Zhuravskaya, and seminar participants at NBER Political Economy Fall Meeting, Harvard, Princeton, Chicago Harris, Bocconi University, Cornell University, 2nd CEPR Workshop on Media, Technology, Politics, and Society, the Winter Meeting Econometric Society, EEA-ESEM conference, University of Cologne, CREST, CESifo, University of Antwerp, University of Bristol, and Toulouse School of Economics for their helpful suggestions. Müller acknowledges financial support from a Singapore MoE start-up grant and Presidential Young Professorship (numbers A-0003319-01-00 and A-0003319-02-00).

†Università Bocconi, Department of Economics, IGIER, Stigler Center, rafael.jimenez@unibocconi.it.
‡National University of Singapore, Department of Finance and Risk Management Institute, kmueller@nus.edu.sg.
§Università Bocconi, Department of Economics, IGIER, PERICLES, CEPR, CAGE, carlo.schwarz@unibocconi.it.

1 Introduction

One of the most frequently voiced charges against social media platforms is that they have amplified existing societal tensions. Forty percent of Americans have experienced some form of online harassment (Anti-Defamation League, 2022), and many are concerned that hateful conversations on social media might contribute to the spread of hateful attitudes offline. Recent empirical evidence suggests that hateful posts on social media can indeed spill over into violent offline actions against ethnic and religious minorities (Bursztyn et al., 2019; Müller and Schwarz, 2021; Müller and Schwarz, 2023).

Social media companies have not sat idle in addressing these problems. Hate speech has been officially prohibited on YouTube since at least 2006, on Facebook since at least 2012, and on 𝕏 (formerly, and henceforth, Twitter) since 2015 (Twitter, 2015; Gillespie, 2018). However, both Facebook and Twitter recently cut back their content moderation efforts, as content moderation remains highly controversial.[1] Some argue that platforms moderate too little, while others worry that moderation leads to censorship and limits the plurality of online discourse. To evaluate whether content moderation policies are socially desirable, it is therefore crucial to understand their online and offline effects.

This paper investigates this question by focusing on the first legal change explicitly aimed at forcing social media platforms to increase their moderation efforts with regard to hate speech: the German "Netzwerkdurchsetzungsgesetz" (Network Enforcement Act, henceforth NetzDG). The NetzDG was enacted on September 1, 2017 in response to a spike in online hate speech during the influx of more than one million refugees into Germany as a result of the 2015-2016 refugee crisis. The law marked an unprecedented legal change that introduced penalties for large social media platforms of up to €50 million for failing to promptly remove hateful content.[2] As such, the law drastically changed social media providers' incentives to moderate such content and has been called a "key test for combatting hate speech on the internet" (Echikson and Knodt, 2022). Since its passing, the NetzDG has effectively become the international standard for addressing illegal online content, and similar legislation has been passed in at least 25 countries, often explicitly referencing the NetzDG as a model (Justitia, 2020).

Despite the widespread adoption of such laws, their efficacy remains an important, unanswered empirical question for at least two reasons. First, it is unclear whether the law

[1] Meta announced in January 2025 that it was terminating its third-party fact-checking program and lifting restrictions on topics that are part of mainstream discourse (Kaplan, 2025), while Elon Musk laid off key staff overseeing content moderation upon acquiring Twitter (Alba and Wagner, 2023).

[2] The NetzDG targeted social media companies with more than two million users. Besides Facebook and Twitter, the law applies or has applied to Change.org, Instagram, Google Plus, YouTube, Pinterest, Reddit, SoundCloud, TikTok, Twitch, and Jodel.

would lead to a lower prevalence of online hate speech, even on targeted platforms, as users may circumvent detection algorithms, or platforms might only use the minimum necessary effort to avoid fines.[3] Second, even if the law effectively curbed hate speech on the targeted online platforms, this does not necessarily translate into lower offline violence if moderation is not sufficiently targeted to radicalized users or if it is implemented too slowly to curb offline harm.

This paper makes progress on both fronts and investigates whether the increased content moderation efforts induced by the NetzDG decreased online and offline hate, and whether content moderation leads to disruptions of online conversations that could reflect an impingement of freedom of speech. Specifically, we focus on toxic online content and real-life hate crimes targeting refugees, given the widespread nature of anti-refugee sentiment in the German context during that period (Müller and Schwarz, 2021). In this environment, the far-right party Alternative for Germany ("Alternative für Deutschland," henceforth AfD) played a crucial role. At the time the NetzDG was enacted, the AfD was the third-largest party in the German parliament, having risen on a platform of anti-refugee rhetoric. Importantly, the AfD had, and still has, the largest Facebook following of any German party. As such, the online activity and geographical spread of AfD Twitter and Facebook followers provide a rich testing ground for analyzing the effects of the NetzDG.

Our empirical analysis proceeds in two parts. In the first part, we focus on the online effects of the NetzDG, including its effects on hatefulness, engagement, and topics of conversation. To measure the impact of the law on the hatefulness of refugee-related Twitter content, we collect the universe of tweets that contain the word "refugee." This data collection allows us to analyze the content of the *surviving* tweets (those not removed by Twitter). We proxy for the hatefulness of these tweets using Google's Perspective API, a machine learning algorithm commonly used in industry applications and as a benchmark in academic studies. This algorithm assigns a "toxicity" score to each tweet, which can roughly be interpreted as the fraction of people who consider it offensive. In a difference-in-differences analysis, we compare the tweets by "toxic users" and "non-toxic users" before and after the NetzDG was implemented. Specifically, we estimate the law's effect by comparing the toxicity of tweets, before and after the legal change, for users who fall into the top vs. bottom quartile of toxicity pre-NetzDG or, alternatively, those who follow vs. not follow the @AfD account on Twitter.

Consistent with an increase in content moderation efforts, we find an immediate and significant decrease in the toxicity of refugee-related tweets after the NetzDG became binding. Compared to the pre-period mean, there is a 28% drop in the toxicity of tweets by users in

[3] Alternatively, platforms may reduce the visibility of hateful content without eliminating it—what Twitter refers to as its "freedom of speech, not reach" philosophy (𝕏 Safety, 2023).

the top quartile of pre-NetzDG toxicity and a drop of around 15% for tweets posted by toxic AfD followers. The results are robust to alternative definitions of toxic users and measures of toxicity, including a measure of threats against an individual or group. We document a similar reduction in the toxicity of overall tweets, suggesting that the NetzDG shifted the nature of online conversations beyond refugee-related content. Given that the Twitter API's historical data provides us with *surviving* tweets, these effects likely reflect both a mechanical removal of hateful tweets and a deterrence effect on the production of such content. Notably, we do not find a decline in refugee-related or overall Twitter activity among toxic users.

Several additional tests suggest that these estimates capture the effect of the NetzDG rather than other concurrent events. The two main potential concerns for our analysis are (i) the end of the refugee crisis and (ii) the 2017 federal election, which are the only major political events that broadly coincide with the introduction of the NetzDG. First, the results cannot be explained by the tailing off of the refugee crisis. We find no evidence that the attention paid to refugees online exhibited a discontinuous shift around the time of the NetzDG. Further, the inflow of refugees into Germany had stopped considerably before the law was passed. Finally, a tailing off of the refugee crisis also cannot explain the decrease in overall online toxicity unrelated to refugee content. Second, we show that the 2017 federal election is also an unlikely confounder. Among others, we provide a placebo check and show that the 2021 federal election was not associated with any changes in online toxicity, suggesting that we are not capturing cyclicality in the hatefulness of online discourse around elections. This test also rules out that our findings are driven by mean reversion in the behavior of toxic users.

Next, we provide additional pieces of descriptive evidence on how online conversations changed around the NetzDG. First, we analyze the frequency of words used by toxic relative to non-toxic users before and after the law. In refugee-related tweets, we find a clear shift away from inflammatory issues such as rape and other forms of sexual violence, terrorism, and Nazi comparisons. Second, we train a machine learning model to identify distinct topics and show that left-leaning topics such as feminism or concerns about antisemitism and neo-Nazis became more prevalent after the NetzDG, while refugee and terrorism-related content received less attention. We find no evidence of changes in discussions of censorship or disengagement with controversial political issues, which suggests that most users did not express concerns that the policy stifled freedom of expression online. Lastly, to gauge the effect on the plurality of discourse on Twitter—which may be reduced as a consequence of censorship (Habibi et al., 2024)—we measure the concentration of topic shares using a Herfindahl–Hirschman Index. We show that there was no change in this measure around the time of passage of the NetzDG, suggesting that the law did not decrease the variety of topics discussed on Twitter.

As a last piece of evidence on the online effects, we study the impact of the NetzDG on the usage of different social media platforms. For this analysis, we collect a panel of web traffic data at the platform-country level, covering Germany and other Western countries as well as several platforms that were not subject to the NetzDG. Using a triple-difference design, we compare changes in the usage of platforms covered by the NetzDG relative to other platforms in Germany as opposed to other countries. These estimates suggest that the NetzDG increased the number of users of the affected platforms. This finding is consistent with the idea that, for many users, platforms that enforce content moderation more stringently might be more attractive, perhaps because hate speech excludes some people from online conversations (Waldron, 2012).[4]

The second part of the paper examines the offline effects of the NetzDG, which are crucial for assessing its overall effectiveness. We investigate whether the policy-induced content moderation efforts by social media platforms translated into fewer real-life hate crimes against refugees. For this analysis, we exploit municipality-level differences in the exposure to far-right social media content. To the extent that the NetzDG limited online hate speech, one would expect a decrease in the number of anti-refugee incidents in areas where more people were exposed to hateful content in the first place. Using a difference-in-differences design, we find that the introduction of the NetzDG led to a reduction of anti-refugee incidents in municipalities with many AfD Facebook followers. Our estimates suggest that the law reduces the number of anti-refugee incidents by 1% for a one standard deviation higher number of AfD followers per capita in a municipality.

The underlying identification assumption of this approach is that in the absence of the NetzDG, municipalities with different prior exposures to hate speech on social media would have seen similar trends in anti-refugee incidents. In support of this assumption, we show that municipalities with different levels of AfD followers had similar trends in hate crimes in the period leading up to the enactment of the NetzDG. Our findings are also robust to controlling for a plethora of municipality characteristics and a large battery of sensitivity checks. For example, the estimates are not driven by differences in support for the AfD in the 2017 federal election, by the level of social media or internet penetration, nor by the number of refugees living in a municipality. These main results remain unchanged if we measure exposure based on Twitter rather than Facebook data, use alternative variable transformations, compute standard errors in various ways, or use more restrictive fixed effects.

We further corroborate our main evidence by using a synthetic control approach, where we compare overall hate crimes in Germany (including those unrelated to refugees) to other

[4]Note that this finding does not imply that it would be profitable for platforms to increase their content moderation efforts. Content moderation—particularly based on user reports (as is the case of the NetzDG)—can be quite resource-intensive, since it requires humans to review content manually (Gillespie, 2018).

countries using a harmonized cross-country dataset. Specifically, we build a synthetic control for Germany using data for the period 2009-2020 from 21 donor countries. Using the full path of pre-intervention hate crimes as predictors, we find that the policy resulted in an annual decrease in 0.03 hate crimes per 10,000 inhabitants, or roughly 250 fewer hate crimes per year. This finding is robust to a battery of robustness checks, including placebo exercises that assign treatment to other countries or that focus on the overall rate of homicides (which should not be affected by the NetzDG).

Finally, we examine two plausible mechanisms for why content moderation prompted by the NetzDG reduced anti-refugee hate crimes. The NetzDG could (i) inhibit collective action targeting refugees, or (ii) change individual attitudes towards refugees. We provide some evidence suggesting that the NetzDG made collective action against refugees more difficult by differentiating incidents by the number of perpetrators. For this analysis, we hand-coded how many persons were involved in an attack on refugees for 10,080 incidents in our data based on a description of each case. In line with the collective action hypothesis, we find that the estimates are twice as large for anti-refugee incidents committed by multiple relative to single perpetrators. This effect is unlikely to be driven by information provision or coordination of attacks online. Rather, the decline in hate crimes may be the result of reduced emotional contagion. These findings are consistent with existing evidence in the literature that collective action may be an important mechanism in explaining the link between online and offline violence (Bursztyn et al., 2019; Müller and Schwarz, 2021).

To study the second channel, we analyze whether the NetzDG led to changes in the attitudes of social media users towards refugees using data from the German Socio-Economic Panel (GSOEP, Goebel et al., 2019). To this end, we investigate within-person changes in attitudes and actions toward refugees, comparing active social media users relative to non-users. We find no evidence of improved attitudes toward refugees, both in the full sample of respondents and a sub-sample of AfD supporters. These findings make it unlikely that more positive attitudes toward refugees are an important explanation for the reduction in anti-refugee incidents.

Overall, our findings suggest that online content moderation can significantly reduce the toxicity of online discourse as well as the prevalence of offline violence. Our evidence also indicates that the NetzDG reduced the use of social media for collective action against refugees, while we find no evidence of major disruptions to discussions of controversial political issues or a reduction in the plurality of topics discussed on Twitter. The increase in the unique number of social media users we document, is also consistent with the idea that content moderation helps to include more users in online conversations. These insights suggest that the NetzDG achieved its primary policy objective, which is of key interest to social media

platforms and policymakers alike. Despite these findings, we want to caution against taking our results as a blanket endorsement of strict content moderation policies, as such policies require a complete welfare analysis to assess the plethora of direct and indirect effects.

Contribution to the literature. The paper contributes to three strands of the literature. First, there is a growing literature on the real-life effects of social media. Existing work has investigated the impact of social media, among other outcomes, on mental health and well-being (Allcott et al., 2020; Braghieri et al., 2022), polarization (Sunstein, 2017; Allcott and Gentzkow, 2017; Boxell et al., 2017; Levy, 2021; Mosquera et al., 2020), protests (Enikolopov et al., 2020; Acemoglu et al., 2017; Fergusson and Molina, 2021; Howard et al., 2011), corruption and confidence in government (Enikolopov et al., 2018; Guriev et al., 2021), and voting (Bond et al., 2012; Jones et al., 2017; Fujiwara et al., 2024). See Zhuravskaya et al. (2020) and Aridor et al. (2024) for reviews of the recent literature on the political effects of social media. Most closely related is the work that provides evidence of the impact of social media on hate crimes (Müller and Schwarz, 2021; Müller and Schwarz, 2023; Bursztyn et al., 2019; Cao et al., 2023) for different social media platforms, countries, and minority groups. However, none of this evidence speaks to platform-induced changes as an effective countermeasure against hateful online content and offline violence, which is of key interest from a policy perspective. Importantly, whether content moderation policies can be effective in reducing online and offline hatred is ex-ante unclear because it depends on how social media users react to increased moderation.

Second, we contribute to a nascent literature that studies platform decisions and content moderation strategies in the context of hate speech and toxic content, theoretically (Liu et al., 2021; Madio and Quinn, 2021; Kominers and Shapiro, 2024; Beknazar-Yuzbashev et al., 2024) and empirically (Beknazar-Yuzbashev et al., 2022; Müller and Schwarz, 2022).[5] Jiménez Durán (2022) finds that moderation has an insignificant effect on consumer surplus, which suggests that the most sizeable welfare effects of content moderation could be due to its impact on out-of-platform outcomes, such as hate crimes. Our findings on online toxicity are consistent with prior work by Andres and Slivko (2021), who estimate the effect of the NetzDG on the toxicity of German vs. Austrian right-wing Twitter users with a difference-in-differences design. In line with our results, they find that German AfD followers posted relatively less toxic content after the NetzDG. Different from their analysis, we focus on *within-country*

[5]There is a parallel literature studying the moderation of misinformation. See Barrera et al. (2020), Henry et al. (2022), and Henry et al. (2023) for recent experimental work comparing different interventions targeting misinformation, and Aridor et al. (2024) for a review. Another related literature is the work on censorship of the internet and social media in autocratic regimes (Qin et al., 2017; Chen and Yang, 2019).

variation for the results on online toxicity. In addition, our paper is the first to jointly study changes in online content, platform usage, hate crimes, and attitudes.

Lastly, we speak to a broader literature on the effects of media on violence. Research by Yanagizawa-Drott (2014), DellaVigna et al. (2014), and Adena et al. (2015), for example, suggests that nationalist propaganda on the radio can increase the prevalence violence against minorities. Djourelova (2023) shows the effect of slanted language on attitudes toward immigrants. In other work, Dahl and DellaVigna (2009), Card and Dahl (2011), and Bhuller et al. (2013) investigate the effect of movies, TV, and the internet on different types of violence. Unlike social media, traditional media undergoes editorial processes and is easier to subject to regulatory oversight. Instead, social media companies indirectly shape content through platform design and content moderation. Nevertheless, our findings suggest that, even in this setting, a policy that imposes penalties (much like a Pigouvian tax) can affect online content and potentially reduce offline externalities.

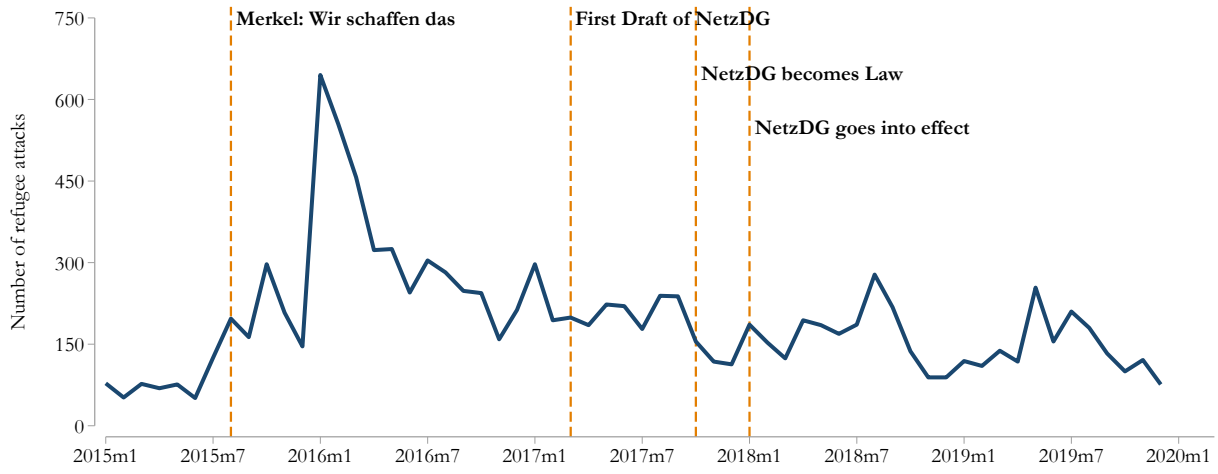
2 Background

In August 2015, Chancellor Angela Merkel declared that Germany would welcome a large number of refugees from the Syrian Civil War and other conflicts who had arrived in Europe in the previous months. Following her "Wir schaffen das!"(we can do this) speech, over 1.3 million refugees entered Germany over the 2015-2016 period. The inflow of refugees only slowed considerably after the European Union struck a deal with Turkey in March 2016, in which Turkey agreed to prevent Syrian refugees from crossing over to the EU in exchange for financial compensation (European Parliament, 2016).

The large inflow of asylum seekers into Germany was accompanied by a flare-up in the number of anti-refugee incidents. The non-profit organization "Amadeu Antonio Stiftung" recorded more than 10,080 hate crimes targeting refugees in Germany between 2016 and 2020, visualized in Figure 1. Hate crimes spiked after Merkel's "Wir schaffen das" speech and peaked following the widely-reported 2016 New Year's Eve sexual assaults by refugees in Cologne. The frequency of these hate crimes also drew the attention of the international news media (see, for example, New York Times, 2017). Importantly, hate crimes against refugees continued even after the flow of refugees to Germany stopped following the EU-Turkey deal in March 2016.

In previous research, Müller and Schwarz (2021) showed that social media played a role in this wave of anti-refugee hate crimes. The Facebook page of the Alternative for Germany (AfD) in particular became an important platform for the spread of anti-refugee content. The

Figure 1: Attacks on Refugees in Germany, 2015-19



Notes: This plot shows the monthly number of refugee attacks in Germany between 2015 and 2019 based on data from the Amadeu Antonio Stiftung, a non-profit organization. The dashed vertical lines mark the date of Merkel's "Wir schaffen das!" speech and important dates in the enactment and implementation of the Netzwerkdurchsetzungsgesetz (NetzDG).

evidence suggests that these far-right Facebook pages helped propagate anti-refugee sentiment, and the exposure to such online content motivated real-world anti-refugee incidents.

In August 2015, Germany's Minister of Justice Heiko Maas demanded that social media companies should enforce existing laws prohibiting hate speech (Economist, 2018). In an open letter, Maas wrote: "The internet is not a lawless space where racist abuse and illegal posts can be allowed to flourish [...]." Due to what he deemed insufficient voluntary action by the social media companies, Maas introduced a first draft of the "Netzwerkdurchsetzungsgesetz" (NetzDG) in March 2017 to stem the wave of hateful content that was circulating on German social media.[6] The first draft of the NetzDG stated explicitly that "hate speech and other criminal content that cannot be effectively combated and prosecuted pose a great threat to peaceful coexistence in a free, open and democratic society" (authors' translation; Deutscher Bundestag, 2017). The NetzDG eventually passed the German parliament on 1 September 2017 and became law on 1 October 2017. Penalties went into effect on January 1st, 2018.

[6]Before the NetzDG, Maas had attempted to work with the major social media companies to reduce the prevalence of hate speech. In December 2015, the Task Force Against Illegal Online Hate Speech—formed by Facebook, Twitter, Google, and some anti-hate advocacy groups in Germany—signed a Code of Conduct. The companies agreed to remove hate speech promptly and to facilitate user reports. However, after several months, Maas noted that "the networks aren't taking the complaints of their own users seriously enough," which led him to introduce legislation with monetary penalties (Kaye, 2019). At the European level, Facebook, Microsoft, Twitter, and YouTube signed a voluntary Code of Conduct with the European Commission in May 2016 to review reported illegal content within 24 hours (Gillespie, 2018). See Gorwa (2019) for a compilation of formal and informal platform governance efforts around that time.

The NetzDG was "the first law that formalizes the process for platform takedown obligations" (Kohl, 2022). Importantly, it did not introduce new definitions of which content was supposed to be illegal. Instead, the NetzDG merely enforced the existing German criminal code, according to which certain types of public expression, such as incitement of criminal activities or genocide, were already illegal, hence the name "Network Enforcement Act." While it was not the first attempt at regulating online content moderation, the law marked a clear shift in the incentives of social media platforms. For the first time, a law established financial penalties of up to €50 million if social media companies with more than 2 million registered users in Germany failed to remove hateful content within 24 hours of being reported by German users.[7]

The first companies to be covered by the law were Google (YouTube and Google+), Meta (Facebook and Instagram), Twitter (now 𝕏), and Change.org (Echikson and Knodt, 2022).[8] To incentivize users to report hateful content, the NetzDG required platforms to implement dedicated buttons to report violations against the law. Appendix Figure A.1 shows an example of such a reporting tool. The law also imposed an unprecedented transparency requirement for platforms to publish a biannual report on their content moderation activities (Heldt, 2019).

While the platforms regulated by the NetzDG do not publish information on the size of their content moderation workforce, there exists anecdotal evidence on the scope of the buildup in content moderation capacities. For example, Facebook already started to build a team of 500 employees specifically tasked to aid compliance with the NetzDG by August 2017, as soon as it became clear the law would pass the German parliament (Heise, 2017). In January 2018, Facebook announced it was employing 10,000 content moderators worldwide and would double that number by the end of the year (Zeit, 2018). Another estimate suggests that Facebook had at least 15,000 content moderators worldwide in 2019, 2,000 of which were based in Germany (Netzpolitik, 2019). Similarly, Twitter announced an expansion in the number of moderators explicitly as a reaction to the NetzDG (Stern, 2018).

The transparency reports that regulated platforms have to publish under the NetzDG offer some insight into the extent of the deletion of illegal content. For example, in 2018 Twitter received around 500,000 complaints and took action in around 10% of these cases

[7] After receiving user reports, companies typically evaluate whether the content violates their guidelines. If so, they take action on the content (e.g., delete a post globally). If the content does not violate their guidelines but comes from a German IP address, it is assessed vis-à-vis the German Criminal Code listed in the NetzDG. If it is considered unlawful under the NetzDG, companies disable access to that content in Germany. See for example, a recent transparency report by Instagram: `https://transparency.fb.com/sr/netzdg-report-english-ig-jul-21`.

[8] Subsequently, other platforms such as Jodel, TikTok, Reddit, SoundCloud, Pinterest, and Twitch started providing the transparency reports required by the law. See `https://www.bundesanzeiger.de/pub/de/suchen2?7`.

(Twitter, 2018a,b). However, the figures on the number of user complaints and the fraction of removed content in these reports is only a lower bound on the true extent of deleted content. To avoid costly penalties, the NetzDG incentivized platforms to adopt stricter proactive measures, such as deleting or deranking hateful content even when users had not flagged it. Such efforts likely expanded significantly alongside the increase in content moderators outlined above. In fact, concerns about such "overblocking" of content have been a major point of contention since the law was first discussed.

Moreover, there are also reasons to believe that the platforms may not always report the numbers accurately in their transparency reports. In 2019, for example, Facebook admitted and was later fined for a severe underreporting of content that users had flagged as being in violation of the NetzDG (Zeit, 2019). Note that the fine was only due to the misreporting of the extent of moderation and not due to Facebook's lag in content moderation.

Overall, the NetzDG marked a pivotal shift in the regulation of online hate speech in Germany. It also provided a template for future legislation regulating hate speech on social media platforms, such as the Online Safety Bill in the United Kingdom and the Digital Services Act of the European Union. By 2020, 25 countries–both democracies and non-democracies–had adopted policies that were often explicitly modeled after the NetzDG (Justitia, 2020). Thus, the NetzDG provides an essential testing ground for the effectiveness of such legislation. In the next section, we describe our main data sources that will allow us to investigate the impact of the NetzDG on online hate speech and offline hate crimes.

3 Data

Our main analysis builds on five separate datasets. First, we construct a database of refugee-related tweets that allows us to study the impact of the NetzDG on the toxicity of online content. Second, we construct a web traffic panel at the country-platform-quarter level that allows us to measure the effect of the law on the user base of treated platforms. Third, for our analysis of the offline effects of the NetzDG, we construct a municipality-quarter panel of anti-refugee incidents. Fourth, we use survey responses from the German Socio-Economic Panel (GSOEP) (Goebel et al., 2019) to study attitudes towards refugees. Fifth, for our synthetic control analysis, we build a cross-country panel of total hate crime. We describe the main data sources for each dataset in the following.

Refugee-related Twitter Content

To provide evidence for the effects of the NetzDG on the toxicity of social media content, we create a tweet-level dataset measuring the online toxicity of refugee-related tweets. We

focus on Twitter data because Facebook, unfortunately, does not allow the collection of posts directly from user profiles. In contrast, Twitter provides rich post and user data, and, importantly, it is also one of the twelve platforms that have been subject to the NetzDG.

We use the full-archive search endpoint of Twitter's Academic API and obtain all tweets containing the word "Flüchtling" (German for *refugee*) between January 2016 and December 2019. As discussed in Section 2, the focus on refugee-related Twitter content is motivated by the increase in online hate speech that occurred during the refugee crisis and the existing evidence that links this online content to offline violence. We thus investigate the effect of the NetzDG on the hatefulness of refugee-related German tweets. In total, this dataset contains 811,332 tweets. Appendix Figure A.2 plots the monthly number of tweets mentioning the word "Flüchtling" (refugee), which shows no downward shift in the number of refugee-related tweets after the implementation of the NetzDG. For the users in our sample, we also collect all other tweets they posted using the Twitter API. In total, this data collection yielded over to 360 Million tweets. This allows us to investigate changes in the overall toxicity of discourse on Twitter. To identify the political leaning of users, we additionally scraped the Twitter follower lists of all major German parties. These lists allow us to identify which Twitter users follow the AfD's Twitter account.

We measure the hatefulness of online content using Google's Perspective API (Wulczyn et al., 2017; Dixon et al., 2018). This API returns a machine-learning-based "toxicity" score between 0 and 1 (where 1 is the most toxic). The score can be interpreted as the fraction of people who consider the content to be "toxic," which is defined as "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion." Besides the main toxicity measure, the API also provides other scores, which include severe toxicity, identity attack, insult, profanity, and threat.[9]

Appendix Table A.2 contains summary statistics for our sample of refugee tweets. On average, refugee-related tweets have a toxicity score equal to 0.33. To get a sense of what kind of language these numbers imply: "Ich mag keine Flüchtlinge" (I don't like refugees) has a toxicity score equal to 0.41, and "Flüchtlinge sind Müll" (Refugees are trash) has a toxicity of 0.8. Around 17% of tweets in the sample were posted by AfD followers. In Appendix Table Table A.1, we provide several examples of toxic refugee tweets in our data.

Platform Usage Panel

To measure the effect of the NetzDG on the usage of websites targeted by the law, we construct a panel dataset with web traffic data at the platform-country-quarter level. We obtain the

[9] See `https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages?language=en_US`.

number of unique users and total visits between January 2017 (the earliest available data) and 2019, for seven major online platforms in 36 OECD countries from Semrush.com.[10] Four of the selected platforms (Instagram, Twitter, YouTube, and Facebook) were the first ones to be subject to the NetzDG in Germany, while three were never subject to it (Amazon, Netflix, and Wikipedia). We focus on the largest platforms because web traffic data is estimated from clickstream data based on a panel of users' browsing behavior (collected from browser extensions and mobile applications), which is imprecise for smaller platforms.[11]

Municipal Anti-Refugee Hate Crime Panel

The analysis of the offline effects of the NetzDG is based on a panel dataset for the number of anti-refugee hate crimes for each German municipality between January 2016 and December 2019, aggregated at the quarterly level. The underlying data on anti-refugee hate crimes were collected by the Amadeu Antonio Foundation and Pro Asyl (a pro-asylum NGO).[12] Information on around three-quarters of these incidents comes from administrative police data reported based on parliamentary requests. All of the 10,080 anti-refugee crimes are classified into four groups. The most common cases are property damage to refugee homes (7,814 incidents), followed by assault (1,693), incidents during anti-refugee protests (72), and arson (153). 348 events are classified as "suspected cases" that are still under investigation. We provide one example for each class of anti-refugee incidents in Appendix Table A.3. We are able to link incidents to their corresponding municipality because they are geo-coded with the exact longitude and latitude. We assign these incidents to municipalities using shape files provided by the ©GeoBasis-DE/BKG 2016 website.[13]

Most of the additional municipality-level variables are based on the replication data from Müller and Schwarz (2021).[14] The main measures of far-right Facebook usage from Müller and Schwarz (2021) we use in our analysis are based on the number of AfD Facebook followers in each municipality, which was obtained by hand-collecting and geo-coding a place of residence for 34,389 users who interacted with AfD's Facebook's page as of October 2017.

[10]These countries are those that joined the OECD before 2020, see `https://www.oecd.org/about/document/ratification-oecd-convention.htm`.

[11]In particular, it is likely that there are not enough observations at the country-quarter-website level for smaller platforms. See `https://www.semrush.com/blog/what-is-clickstream-data/`. For example, Change.org, which was the one other platform among those initially subject to the NetzDG, has only 1.6% of Facebook's traffic, according to Semrush estimates.

[12]These data are available at `https://www.mut-gegen-rechte-gewalt.de/service/chronik-vorfaelle`.

[13]The analysis is conducted on the level of 4,679 German municipalities ("Gemeindeverwaltungsverband"). After removing uninhabited areas, we are left with 4,466 municipalities in our sample. We use the level of the "Gemeindeverwaltungsverband" instead of "Gemeinden" since the area and population of these administrative areas are more similar.

[14]The underlying reproduction file is available here.

We base our exposure measure on the AfD's Facebook page because the right-wing populist party's Facebook page was arguably the key platform for anti-refugee content online during the period we study and has a broader reach than the Facebook page of any other German party. Moreover, we focus on Facebook because it is the most widely used platform in the German setting. We augment these data with information about the activity of each user, which allows us to construct the number of posts, likes, comments, and shares for each AfD user.[15]

We visualize the relationship between far-right Facebook usage and hate crimes in Figure 2. The map overlays quintiles of AfD Facebook usage per capita overlayed with the location of anti-refugee incidents (orange dots). There is considerable geographical variation in both incidents and AfD users. Appendix Table A.4 presents summary statistics for anti-refugee incidents, our measure of exposure to online hate speech (AfD users per capita), and our control variables. The unit of analysis is a municipality-quarter. There was at least one incident in every quarter of our study period, and 48% of municipalities experienced at least one incident. On average, municipalities have 3 AfD users per 10,000 inhabitants and 80% have at least one AfD user.

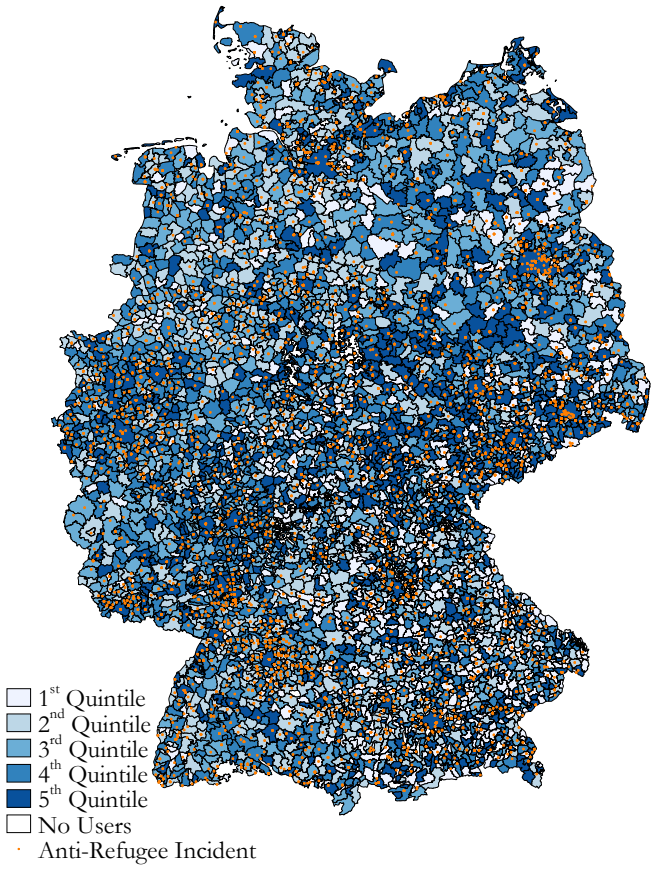
To control for the number of Facebook users in a municipality, we create a measure using Google search. In particular, we use a list of the names of over 2,000 German cities as well as all German municipalities and use the Google Search API to obtain the number of people who indicate living in each municipality on their Facebook profile. To do so, we search for "Lives in: *City Name*" restricted to Facebook.com, where *City Name* corresponds to either a city's or municipality's name. These Google searches return the number of Facebook user profiles where people indicate living in a particular municipality, which should be a sound proxy for the number of local Facebook users.

For robustness, we construct an alternative exposure measure based on the number of AfD followers on Twitter in a municipality. For this measure, we use the location information from the user profiles that we collected for our analysis of Twitter content. This data allows us to verify our findings based on the exposure to hateful content on an alternative social media platform.

Finally, we add municipality-level controls for socioeconomic factors and measures of voting and media consumption behavior. The main source of socioeconomic data is the German Statistical Office, which disseminates regional data via www.regionalstatistik.de. For each municipality, we can measure population by age group, GDP per worker, population density, and the vote results for the German Federal Election in September 2017. We also have

[15]The number of shares of each AfD follower on Facebook was not included in the replication file but stem from the same data collection effort.

Figure 2: Map AfD Facebook Users and Anti-Refugee Incidents



Notes: The shading of the maps indicates the quintiles of the distribution of AfD users per capita for the municipalities in Germany. Orange dot indicate anti-refugee incidents.

data on the share of immigrants and asylum seekers. Data on broadband internet availability comes from the Federal Ministry of Transport and Digital Infrastructure (BMVI). To measure the popularity of traditional media, we use data for 2016-2017 newspaper sales from the "Zeitungsmarktforschung Gesellschaft der deutschen Zeitungen (ZMG)" (Society for Market Research of German Newspapers), which we normalize by population. Data on other types of crimes, which are reported by county and year, come from the Bundeskriminalamt (BKA)'s Police Crime Statistics.

Survey Data on Attitudes Towards Refugees

To study whether the NetzDG was associated with changes in attitudes towards refugees, we extract a set of relevant questions from the German Socio-Economic Panel (GSOEP) (Goebel et al., 2019). The GSOEP is the largest yearly household panel survey in Germany. In the 2016 and 2018 waves of the GSOEP, respondents were asked a battery of questions about their attitudes toward refugees. For example, respondents were asked if refugees are good for the economy, for culture, or for their place of residence. Furthermore, the GSOEP asks whether respondents have taken actions to help refugees (e.g., by donating or volunteering).

We harmonize the coding of all questions into indicator variables such that 1 represents a positive attitude or action towards refugees. We additionally created indices that capture the average response of a respondent across the attitude and action questions. Further, we use questions on social media habits to create an indicator for respondents who use social media at least once a week. We provide summary statistics for the full set of questions and derived variables in Appendix Table A.6.

Cross-Country Hate Crime Panel

We construct a cross-country panel of hate crime incidents for the years 2009-2020, which enables us to construct a synthetic control for Germany. The most comprehensive hate crime database covering several countries is compiled by the Organization for Security and Co-operation in Europe (OSCE). We obtained the reported hate crimes for each of the 57 member States of the OSCE as well as meta data describing measurement changes over time.[16]

The data that Germany reports to the OSCE, however, include online hate speech offenses. To avoid picking up a spurious effect from changes in hate speech reporting due to the NetzDG, we obtain data on violent hate crimes (which do not include hate speech) from Germany's Federal Ministry of the Interior and Homeland (BMI) from the table *Übersicht "Hasskriminalität:" Entwicklung der Fallzahlen 2001 – 2021.* Violent hate crimes include bomb attacks, arson attacks, homicides (including attempts), robberies, physical injuries, and violent property damages. Lastly, we gathered population counts from the World Bank's World Development Indicators.

Table A.7 summarizes the data availability for the OSCE members and the filters that we impose in order to build a balanced panel of countries, which we describe in more detail in Appendix A. We excluded micro-states, countries that changed their measurement of hate crimes after the NetzDG, and countries with more than 50% (six) missing observations in

[16]The underlying data can be downloaded from `https://hatecrime.osce.org/country`. The information on reporting changes is available `https://hatecrime.osce.org/national-frameworks-country#dataCollection`.

2009-2020. To retain as many countries as possible, we linearly interpolate the gaps for the remaining countries but discard those with missing values at the beginning or end of the series. The resulting dataset contains 21 countries in addition to Germany. Appendix Figure D.3 shows the evolution of hate crimes in Germany and the raw mean of the donor countries. Unsurprisingly, we find that the large differences in pre-existing trends across countries make a traditional differences-in-differences analysis difficult.

4 Online Effects of the NetzDG

In the first part of the paper, we investigate the online effects of the NetzDG. The analysis proceeds in three steps. We start by studying the impact on the toxicity of social media content, particularly when it is related to refugees. Then, we investigate whether the NetzDG affected content, such as word frequencies and topics, among other dimensions. Finally, we analyze the effects of the law on platform usage.

4.1 Impact of the NetzDG on Online Toxicity

As outlined in Section 2, the NetzDG marked a clear shift in the incentives of social media companies to moderate online content. In Online Appendix B, we provide a theoretical framework to derive predictions on how such a change in incentives may affect the prevalence of hateful content. In our framework, we interpret the NetzDG as a tax that increases the marginal cost of the prevalence of unmoderated hate speech on social media platforms. In the context of a dominant platform—such as Facebook in Germany, which had a 95% market share of daily active users in 2018 (Bundeskartellamt, 2019)—the framework predicts that this policy should result in a decrease in the equilibrium amount of unmoderated hate speech on the platform.

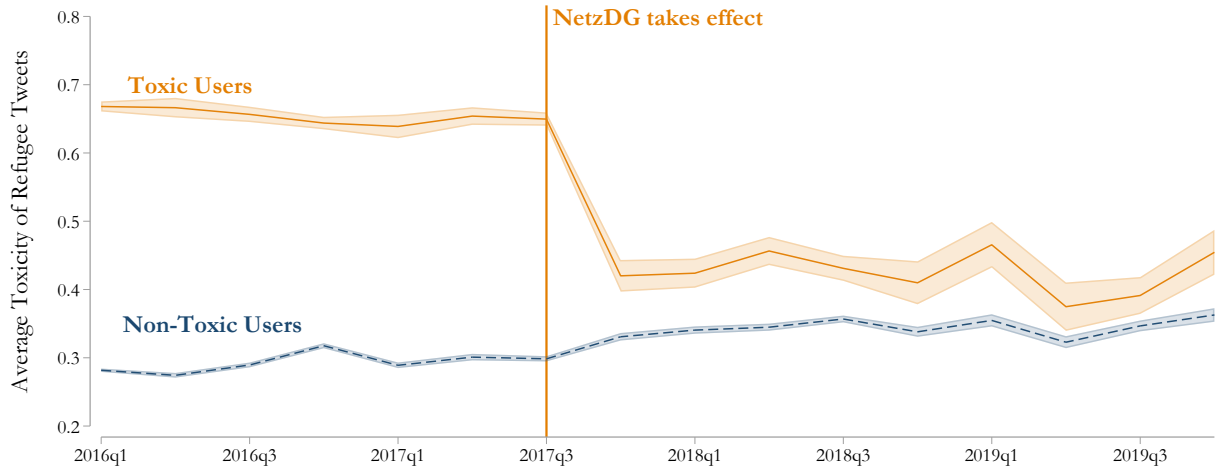
Empirical Strategy

Our strategy compares changes in the toxicity of refugee-related tweets by users producing particularly toxic content to other Twitter users, before and after the implementation of the NetzDG. In particular, we expect to see a decrease in the average toxicity of refugee-related tweets that survive on the platform by more "exposed" users relative to others. We compare toxicity before and after the NetzDG even if, technically, the law could also affect content that was posted before its implementation. However, since the NetzDG relied heavily on users flagging content, newly posted content was more likely to be reported, as it featured more prominently in users' timelines. As a result, the NetzDG disproportionately affected

platforms' incentive to delete or hide content posted *after* it went into effect. Note that any content moderation of social media content that was posted before the NetzDG would bias our results towards 0. Our estimates are, therefore, likely a lower bound.

To motivate our empirical analysis, Figure 3 plots the average toxicity of "surviving" tweets by "toxic" and "non-toxic" users around the passage of the NetzDG. We classify users as toxic if their tweets were, on average, in the top decile of the toxicity distribution before the NetzDG.[17] The figure reveals a striking pattern: while the toxicity content of both groups of users was more or less constant in the lead-up to the NetzDG, there is a large drop in the toxicity of toxic users after the passage of the NetzDG that persists until the end of our observation period.

Figure 3: Toxicity of Refugee Tweets before and after the NetzDG



Notes: 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 were on average above the 90th percentile of the toxicity distribution. "Non-Toxic Users" are all remaining users. The shaded areas indicate 95% confidence intervals for the means.

Motivated by these time series patterns, we estimate a difference-in-differences regression of the following form:

$$Toxicity_{it} = \theta \cdot Toxic\ User_i \times Post\ NetzDG_t + \phi_i + \mu_t + \psi_{it}, \tag{1}$$

where $Toxicity_{it}$ denotes the average toxicity score of tweets by user i in quarter t, based on the coding from the Google Perspective API. The main independent variable is the interaction

[17] We focus on the upper tail of the toxicity distribution as these users were the most likely target of content moderation. In our later results, we show that the choice of cutoff is immaterial to our findings.

between our exposure measure—an indicator variable for highly toxic users ($Toxic\ User_i$)—and the post-NetzDG dummy ($Post\ NetzDG_t$). $Post\ NetzDG_t$ is equal to 1 after the third quarter of 2017 when the NetzDG was enacted (on 1 September), and the platforms had started hiring content moderators (see Section 2).

We show results for two definitions of $Toxic\ User_i$. One version defines exposed users as those that sent particularly toxic content before the NetzDG, defined as those with an average pre-period toxicity of above the 75$^{\text{th}}$ percentile.[18] As a second definition of $Toxic\ User_i$, we take highly toxic users who follow the AfD, motivated by the fact that the AfD positioned itself as a clear anti-refugee voice in Germany (Müller and Schwarz, 2021). To prevent the estimates from picking up shifts in user composition, we restrict this analysis to users who were active in the pre-period and joined Twitter before January 2016. The difference-in-difference regressions also allow us to control for user and time-specific factors with a full set of fixed effects. We cluster standard errors at the user level.

Results

Table 1 presents the results from estimating equation (1). Columns (1) and (2) show the results for users who posted highly toxic content before the NetzDG. Columns (3) and (4) show the results for AfD followers. All specifications indicate a significant reduction in the toxicity of tweets after the NetzDG. The results hardly change when we include user-specific linear time trends in columns (2) and (4). The estimates for highly toxic users in column (2) suggest that the NetzDG was associated with a reduction in the toxicity of tweets of around 28% relative to the pre-period mean of the toxic users (mean=0.528). To provide a more intuitive understanding for the coefficient of -0.145, this is approximately the difference in the toxicity of the statements "Der Flüchtlingsabschaum muss mit Gewalt aus Deutschland vertrieben werden, er ist ein krimineller" (refugee scum must be violently removed from Germany, they are criminals), with a toxicity score of 0.72, and the statement "Diese Flüchtlinge müssen aus Deutschland ausgewiesen werden, sie sind Kriminelle" (these refugees must be removed from Germany, they are criminals) with a toxicity score of 0.57. The magnitude for tweets by AfD users in column (4) is 15% relative to the pre-period mean of toxic AfD users (mean=0.502).[19] These results suggest that the reduction in toxicity is slightly smaller for tweets posted by users with a stronger ideological attachment to the AfD.

[18] In Appendix Table C.1, we show that our results hold irrespective of the precise cutoff. Note that to be included in the analysis, a user needs to have posted at least one refugee-related tweet before the NetzDG, as we are otherwise unable to calculate the pre-period toxicity.

[19] Andres and Slivko (2021) find a reduction of around 2.5% in the monthly volume of hateful tweets about migration and religion sent in Germany relative to Austria.

Table 1: NetzDG and Refugee-Related Online Toxicity

	Dep. var.: Average Toxicity of Refugee Tweets			
	(1)	(2)	(3)	(4)
Toxic User × Post	-0.166***	-0.145***		
	(0.004)	(0.009)		
Toxic AfD User × Post			-0.104***	-0.075***
			(0.009)	(0.021)
User FE	Yes	Yes	Yes	Yes
Quarter FE	Yes	Yes	Yes	Yes
Linear Time Trend		Yes		Yes
Observations	86,208	86,208	86,208	86,208
Pre-Period Mean of DV	0.31	0.31	0.31	0.31
R^2	0.44	0.63	0.43	0.63

Notes: This table presents the results of estimating Equation (1), where the dependent variable is the toxicity of tweets containing the word "Flüchtling" (refugee) (bounded between 0 and 1). In columns (1) and (2) $Toxic\,User_i$ is an indicator variable equal to 1 if a users' refugee tweets before the NetzDG were on average above the 75th percentile of the toxicity distribution. In columns (3) and (4), $Toxic\,User_i$ is an indicator variable that is equal to 1 if a Twitter user additionally follows the AfD's account. All regressions control for user and quarter fixed effects. Columns (2) and (4) additionally control for user-specific linear time trends. Robust standard errors in parentheses are clustered by user. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

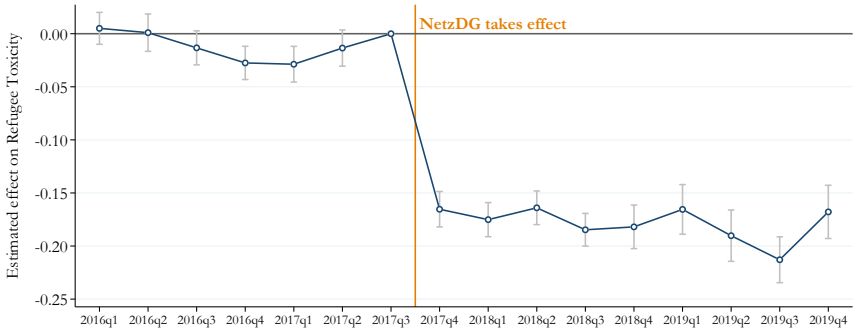
Figure 4 shows a dynamic event study version of these specifications, which replaces the *Post* indicator variable with dummies for the quarters around when the NetzDG became binding. Panel (a) shows the event study for highly toxic Twitter users, and Panel (b) that for AfD followers. Both figures suggest that the refugee-related tweets by highly toxic users and AfD followers were on similar trends of toxicity compared to other Twitter users up to 2017q3, when the NetzDG was enacted. As soon as the law took effect in 2017q4, toxicity quickly and persistently decreased, and consistently remained below its pre-period level. The timing of the effect is also in line with the hiring of content moderators by social media platforms (see Section 2).

Overall Online Toxicity

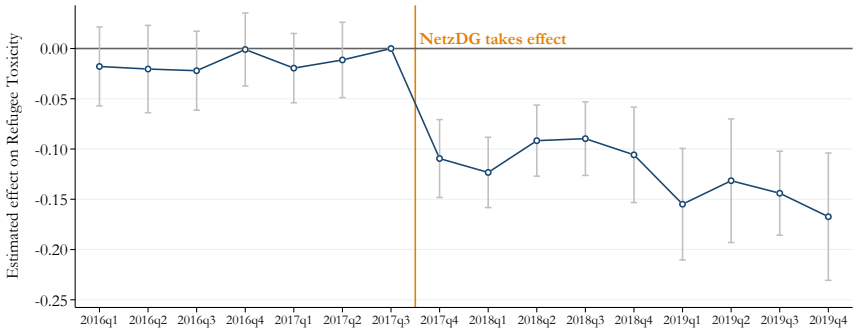
Appendix Figure C.2 provides additional evidence on the effect of the NetzDG by investigating the toxicity of all Twitter content, that is, without restricting the sample to refugee-related tweets. The figure plots an event study for the average toxicity of tweets sent by highly toxic users. Similar to the results for refugee-related content, we observe a significant reduction in overall online toxicity after the NetzDG. The regression estimates from this analysis are shown in Appendix Table C.5. These findings also allow us to address concerns that toxic

Figure 4: NetzDG and Online Toxicity of Refugee-related Content

(a) Toxic Users



(b) Toxic AfD Followers



Notes: Panels A and B plot the coefficients from event study versions of Equation (1). In Panel (a), we define $Toxic\,User_i$ equal to 1 if a user was in the top quartile of refugee toxicity pre-NetzDG, and 0 otherwise. In Panel (b), we define $Toxic\,User_i$ equal to 1 if a user was in the top quartile of refugee toxicity pre-NetzDG and followed the AfD. The dependent variable is the toxicity of tweets containing the word refugee ("Flüchtling"). The omitted category is the 3rd quarter of 2017, the quarter when the NetzDG was enacted (on 1 September) but before it took effect (on 1 October), as indicated with the vertical line. The whiskers indicate 95% confidence intervals based on standard errors clustered by user.

users might attempt to circumvent the NetzDG by substituting away from refugee-related posts (although it should be noted that we do not find any decrease in the frequency of refugee tweets after the NetzDG; see Appendix Table C.4).

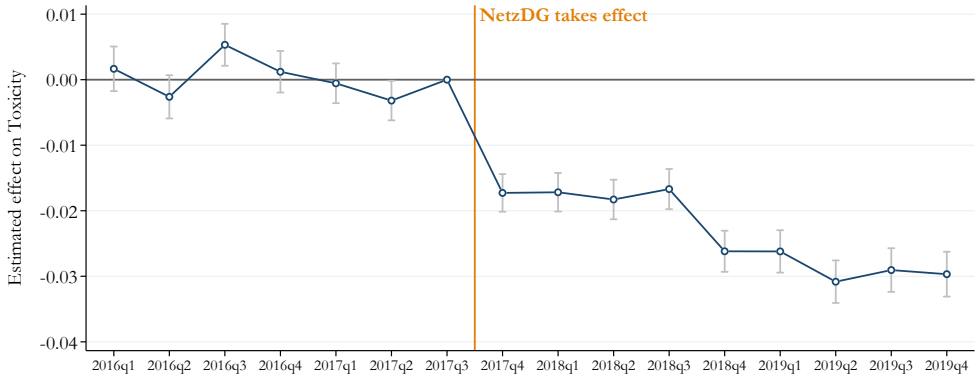
Alternative Explanations

As in any difference in difference-in-differences strategy, we require that no other shock affects toxic users at the same time as the NetzDG. While the sharp timing of the drop in online toxicity allows us to rule out many other potential confounding events, there are two potential concerns regarding our interpretation that the observed reductions in online toxicity were driven by the NetzDG. First, the passage of the NetzDG may coincide with the end of the

refugee crisis, which may itself be associated with reduced refugee-related toxicity. Second, the passage of the NetzDG coincides with the 2017 federal election in Germany. We address these concerns in the following.

The hypothesis that our findings are driven by the refugee crisis in Germany strikes us as unlikely for three reasons. First, the end of the refugee crisis cannot easily explain why we observe an overall drop in the toxicity of online content. This drop is still visible when we exclude refugee-related Twitter content (see Figure 5). Second, the NetzDG did not overlap with an important demarcation point for the refugee crisis in Germany. For example, the deal with Turkey that stopped the inflow of refugees had been struck more than a year earlier (see Section 2). Third, we also do not observe a shift in the attention paid to the topic of refugees in Germany as measured by either by the number of tweets (see Figure A.2) or by the number of Google searches on the topic (see Appendix Figure C.1). There is also no evidence of reduced activity by toxic users (see Table C.4 and Table C.9). This stands in stark contrast to the immediate and large drop in toxicity we observe. There was also no other coinciding political event that should have affected online toxicity or the stance towards refugees. Taken together, these findings make it highly unlikely that our findings merely reflect the end of the refugee crisis in Germany.

Figure 5: NetzDG and Overall Online Toxicity without Refugee Tweets

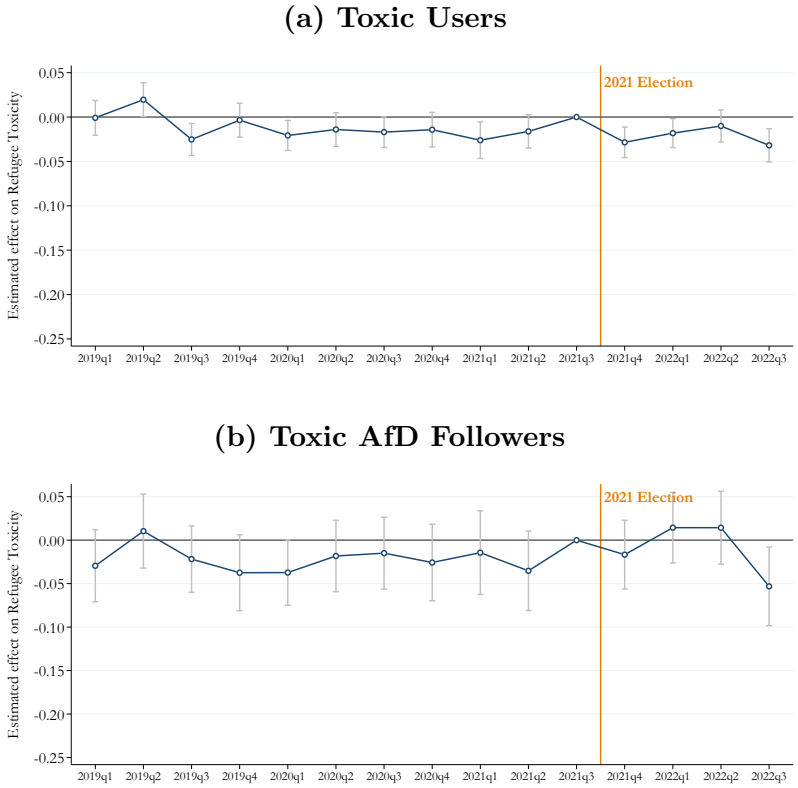


Notes: The figure plots the coefficients from event study versions of Equation (1). The dependent variable is the average toxicity of all tweets (excluding refugee tweets) sent by the users from our main analysis. The omitted category is the 3rd quarter of 2017, the quarter before the passing of the NetzDG (indicated with the vertical line). The whiskers indicate 95% confidence intervals based on standard errors clustered by user.

An alternative hypothesis could be that the drop in toxicity is the result of the end of the 2017 election cycle in Germany. This hypothesis also strikes us as unlikely for several reasons. First, the time series of the average toxicity in Figure 3 shows no increase in toxicity during the election period, suggesting that there is no election cycle in toxicity in Germany during that time. Second, based on the existing evidence of far-right electoral success (e.g., Bursztyn

et al., 2020; Müller and Schwarz, 2023), the strong showing of the AfD in the 2017 election should, if anything, encourage supporters to produce more, not less toxic messages. Third, our data allow us to investigate the evolution of toxicity around the 2021 federal election in Germany (see Figure 6), which was not followed by a drop in toxicity. This placebo test also let us rule out that the mean reversion in the behavior of toxic users might drive our findings. If toxic users mechanically became less toxic over time, we would see the same in this placebo check. Overall, we conclude that our estimates are most likely to reflect the causal effect of the NetzDG.

Figure 6: 2021 Election and Online Toxicity of Refugee-Related Content



Notes: Panels A and B plot the coefficients from event study versions of Equation (1). In Panel (a), we define $Toxic\,User_i$ equal to 1 if a user was in the top quartile of refugee toxicity pre-2021 Election, and 0 otherwise. In Panel (b), we define $Toxic\,User_i$ equal to 1 if a user was in the top quartile of refugee toxicity pre-2021 Election and followed the AfD. The dependent variable is the toxicity of tweets containing the word refugee ("Flüchtling"). The omitted category is the 3rd quarter of 2017, the quarter before the passing of the NetzDG (indicated with the vertical line). The whiskers indicate 95% confidence intervals based on standard errors clustered by user.

Robustness

We conduct several robustness checks to validate our findings. In Appendix Table C.1, we consider different cutoffs of pre-period toxicity for defining "treated" users. Across all specifications, we find a reduction in online toxicity after the passing of the NetzDG. Appendix Table C.2 presents robustness exercises using different measures of toxicity produced by Google's Perspective API. The effect is consistently significant and negative across almost all toxicity measures. Appendix Table C.4 presents estimates of the impact of the NetzDG on the overall amount of refugee-related content that users produce. For both highly toxic users and AfD followers, the number of refugee-related tweets did not decrease after the passage of the NetzDG.

We also conduct the same robustness checks for the results based on all tweets (see Appendix Table C.6). The findings hold independent of the toxicity measure we use (see Appendix Table C.7). Notably, column (6) of Table C.7 suggests a negative effect of the NetzDG on a measure of the threatening language of tweets, which may suggest a role of content moderation in obstructing the collective sentiment underlying violent acts.[20] We will revisit this hypothesis in Section 5.2. We further show that the overall Twitter activity of toxic users, if anything, increased after the NetzDG (see Appendix Table C.9).

Another concern with our analysis is that technically all users, and therefore also our control group, are treated by the NetzDG. While this would not invalidate our empirical strategy, it could pose a challenge to the interpretation of our estimates. As we observe hardly any shifts in toxicity of non-toxic users in the time series (see Figure 3), this seems to be less of an issue in our setting. We additionally tried alternative specifications in which we define toxic users not based on their average but their maximum toxicity, e.g., users who posted any tweet with a toxicity above 0.5 before the NetzDG. The results from this exercise are shown in Appendix Table C.3. The patterns are remarkably similar to our baseline estimates and even hold if we use a threshold of 0.2, in which the control group only contains users who tweet low-toxicity content.

Interpretation

The collection of evidence presented above suggests that the NetzDG induced a reduction in the hatefulness of online content. This reduction is likely driven by a combination of three factors. First, online platforms significantly increased their moderation efforts after the NetzDG. Besides the direct removal of content, platforms could have adjusted their algorithms

[20] The Perspective API defines threats as "Describes an intention to inflict pain, injury, or violence against an individual or group." See `https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages`.

to reduce the exposure of German users to hateful content, partly in anticipation of legal concerns. Second, as toxic tweets provoke further toxic tweets (e.g., Müller and Schwarz, 2023; Müller and Schwarz, 2022; Beknazar-Yuzbashev et al., 2022), content moderation may have a multiplier effect by which the removal of toxic content prevents additional toxic tweets downstream. Third, the NetzDG could have deterred users from posting toxic content in the first place by affecting their first or second-order beliefs. For example, users may have become concerned about the potential legal repercussions of posting toxic messages (even though actual legal cases are extremely rare). Alternatively, the NetzDG could have changed users' second-order beliefs about how acceptable other users find toxic content.

Given that these channels interact with each other in equilibrium, it is impossible to disentangle their individual contribution to the aggregate effect we document. Importantly, all three of these mechanisms are in line with the interpretation that the NetzDG was effective in reducing the toxicity of social media content, which is what matters for our analysis. This drop in toxicity motivates the analyses in the second part of the paper, where we examine whether the NetzDG-induced reduction in hateful online rhetoric also affected real-life anti-refugee incidents.

The absence of any decline in online activity among toxic users further suggests that the NetzDG did not alter their revealed preference for the platform or prompt them to leave it altogether.

4.2 Changes in Online Content

As the second step of our analysis, we investigate other changes in the online discourse besides measures of toxicity. In particular, we study changes in *what* gets discussed instead of only focusing on *how*. In line with our previous analysis, we focus on the period from January 2016 to December 2019. Note that the goal of this analysis is not to establish the causal effect of the NetzDG on the topics of online discourse, but rather to provide descriptive evidence for changes in the topics of online debate around the NetzDG.

We use two approaches. First, we study changes in the frequency of words used by toxic relative to non-toxic Twitter users in refugee-related tweets. Second, we use a machine learning model to classify distinct sets of topics and analyze overall shifts in the issues that are being discussed online.

Changes in Word Frequencies

As the first test, we analyze changes in the word frequencies of refugee-related Twitter content for toxic relative to non-toxic users before and after the NetzDG. Let p_{wgt} be the probability

of word w being used by group $g \in \{0, 1\}$ (non-toxic (0) or toxic (1) user) in period $t \in \{0, 1\}$ (before (0) or after (1) the NetzDG). We calculate:

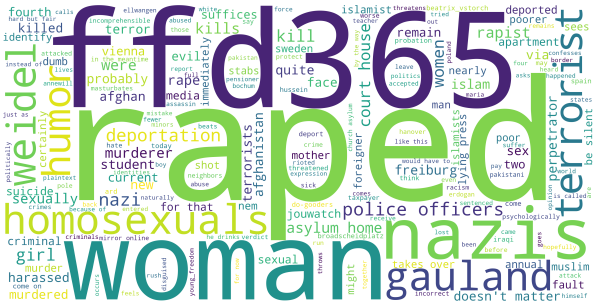
$$\Delta_w = (p_{w11} - p_{w10}) - (p_{w01} - p_{w00})$$

Put differently, we calculate the change in word frequencies for toxic and non-toxic users after the NetzDG. Then, we take the difference in these changes between toxic and non-toxic users. This exercise allows us to understand which words saw the greatest changes in usage among toxic users relative to other Twitter users.[21]

Figure 7 visualizes the results from this analysis. Panel (a) shows words with decreasing and Panel (b) those with increasing frequency. For convenience, we translated everything into English. From this analysis, three findings stand out. First, there is a clear shift away from inflammatory issues such as rape and other forms of sexual violence, terrorism, and Nazi comparisons. The prominently decreasing term "ffd365" was a now-defunct right-wing news website. Second, there is an increase in mentions of mainstream political actors (Merkel, CDU, AfD) and words that suggest a more nuanced debate on the topics, mentioning difficulties, integration, and traumatization. Third, we observe a minor increase in the mentions of "freedom of speech." This could hint at a small number of users being concerned about restrictions to online content, even though we do not observe increased direct discussions about the NetzDG and censorship.

Figure 7: Changes in Word Frequencies – Refugee Tweets

(a) Decreasing Frequency



(b) Increasing Frequency

Notes: This figure shows word clouds that visualize the relative word frequency changes for toxic compared to non-toxic users after the NetzDG in a sample of refugee-related tweets. Panel (a) shows words with decreasing relative frequency, while panel (b) shows words with increasing relative frequency. The size of the words is proportional to the frequency change.

[21] For the calculation of word frequencies, we cast all words in lowercase, exclude stopwords (very frequent words), and restrict our analysis to the 1,000 most frequent words.

Appendix Figure C.3 displays the word frequency changes for all tweets without restricting to refugee-related content. Panel (a) again shows the words with decreasing frequency, and panel (b) shows the words with increasing frequency. There is an overall shift away from political topics and refugee-related issues in particular. There are also fewer mentions of political leaders (e.g., Angela Merkel, Erdogan), political organizations (e.g., AfD, CDU, Pegida), and refugee-related terms (e.g., refugee, Islam, Muslims, terror). Lastly, we also observe a shift from mainstream news outlets (e.g., Welt, Spiegel Online, Zeit Online, NTV) to video platforms like YouTube.

Changes in Topics

As a second analysis, we investigate overall topic changes around the NetzDG using machine learning models that identify individual topics. Topic models describe a range of techniques that make it possible to automatically group similar text documents into topics. Each topic, in turn, is described by a set of frequent topic words.[22] We use the *top2vec* method (Angelov, 2020), which combines pre-trained semantic embedding—a technique to represent text as vectors with low dimensionality—with clustering algorithms to identify topics. This model has at least three crucial advantages for our setting. First, the use of pre-trained semantic embeddings allows the model to use information from vast outside corpora to infer the relationships between words, which is particularly helpful for short texts such as tweets. Second, top2vec automatically finds the number of topics instead of us having to choose a topic in an ad-hoc manner. Third, top2vec is able to infer far more finely-grained topics than other commonly used methods such as Latent Dirichlet Allocation.[23]

As the training of topic models is computationally intensive, we restrict this analysis to a random subset of one million tweets, which nevertheless should suffice to accurately capture overall topic dynamics. As a preliminary step, we remove links and mentions of accounts from the tweets, as well as the hashtag sign ("#"). This procedure ensures, for example, that "#refugee" is treated equally to the word "refugee." We then fit top2vec to this corpus using the "distiluse-base-multilingual-cased" embeddings model (Reimers and Gurevych, 2019) and "hdbscan" (Campello et al., 2013) as a clustering algorithm. To ensure that the topics are meaningful, we additionally specify that the model creates clusters with at least 250 tweets. The resulting model creates 278 topics. For the analysis, we then calculate the share of each

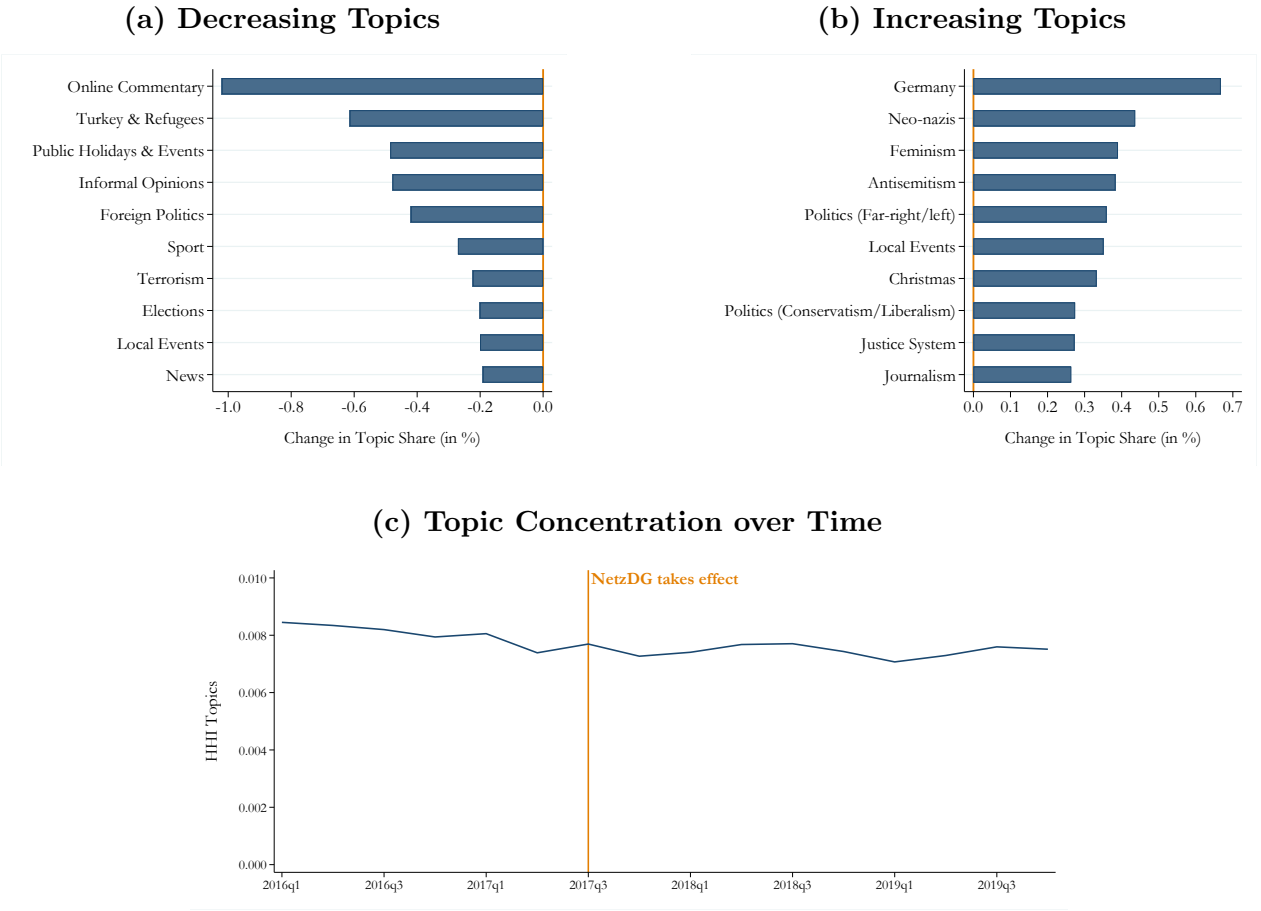
[22] In recent years, topic modeling techniques have found countless applications in the social sciences and have, among many others, been used to analyze journal articles (Griffiths and Steyvers, 2004), transcripts of the Federal Reserve's Open Market Committee (Hansen et al., 2018), the history of economic thought (Ambrosino et al., 2018), ideologies (Draca and Schwarz, 2024), and parenting styles (Rauh and Renée, 2023). See Gentzkow et al. (2019) and Ash and Hansen (2023) for more details.

[23] Latent Dirichlet Allocation is the most widely used topic model (Blei et al., 2003). See Schwarz (2023) for a Stata implementation.

topic among all tweets for the period from 2016q1 to 2017q3 (pre-period) and for the period from 2017q4-2019q4 (post-period).

The results from this analysis are shown in Figure 8. Panel (a) shows the topics with the largest decrease in their topic share, and panel (b) shows the topics with the largest increase. The y-axis contains the topic labels we manually assigned based on the topic words and a reading of the tweets. Appendix Table C.10 shows the full list of words for each of the 20 topics.

Figure 8: Changes in Topics – All Tweets



Notes: This figure shows the ten topics with the largest relative decrease (panel a) or increase (panel b) in their topic share after the NetzDG. The topics were created using the top2vec (Angelov, 2020) topic model. The y-axis lists the five most important topic words for each topic. In panel (c), we plot the Herfindahl-Hirshman-Index (HHI) as a measure of the concentration of topic shares over time.

The topic model suggests an increased discussion of left-leaning topics like antisemitism, feminism, and concerns about neo-Nazis. We also see more debate around Germany in general and German Turks in particular. Topics of decreasing importance are, among others, Turkey and refugees, terrorism, and foreign policy. The first two topics, in particular, are important

electoral issues for the Alternative for Germany. Overall, the topic model results are consistent with a shift of Twitter discussion towards somewhat more left-leaning topics in the German context. In line with the results based on word frequencies, we do not observe a strong rise in discussions of censorship. The results also do not suggest that people disengage from controversial political issues.

In Panel (c), we visualize the concentration of topic shares over time using a Herfindahl–Hirschman Index (HHI). The topic concentration measure is given by:

$$HHI_t = \sum_{i=1}^{T} s_{it}$$

where s_{it} is the share of topic i in quarter t. Overall, the HHI exhibits a slight downward trend, indicating a marginal decrease in topic concentration. Moreover, we observe no significant changes around the passage of the NetzDG. This exercise suggests that the NetzDG did not result in a disproportionate removal of certain contentious topics—as such a shift would have increased the HHI measure by concentrating the remaining content on fewer topics.

4.3 Impact of the NetzDG on Platform Usage

As the last step of our analysis on the online effects of the NetzDG, we investigate its impact on platform usage. One major concern with the NetzDG was that it could stifle the usage of the moderated platforms. While we found no reduction in the number of tweets of toxic users relative to non-toxic users after the NetzDG (see Appendix Table C.4 and Table C.9), there could nonetheless be significant changes in the overall usage of moderated platforms.

We investigate this possibility based on web traffic data from seven major online platforms in 36 OECD countries provided by Semrush. By "moderated," we mean the four platforms initially subject to the NetzDG in Germany (Instagram, Twitter, YouTube, and Facebook), while by "unmoderated" we mean three that were not (Amazon, Netflix, and Wikipedia). Equipped with these data, we estimate the following triple-difference regression:

$$\begin{aligned} Usage_{ict} = {} & \beta_1 \cdot Moderated\ Platform_i \times Germany_c \times Post\ NetzDG_t \\ & + \beta_2 \cdot Moderated\ Platform_i \times Post\ NetzDG_t \\ & + \beta_3 \cdot Germany_c \times Post\ NetzDG_t \\ & + \gamma_i + \omega_c + \delta_t + \epsilon_{ict}, \end{aligned} \tag{2}$$

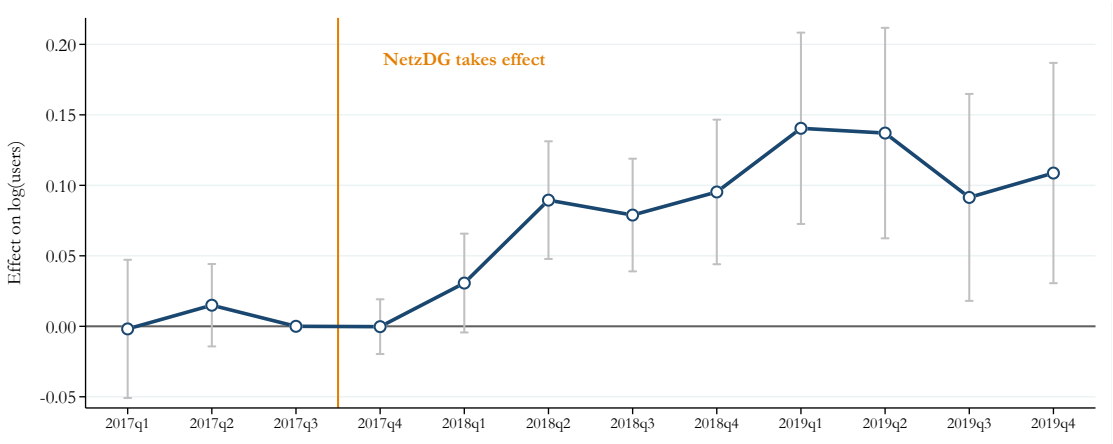
where $Usage_{ict}$ is either the log number of users (unique visitors) or the log number of total visits of platform i in country c in quarter t. $Moderated\ Platform_i$ is an indicator of whether

the platform is subject to the NetzDG in Germany, $Germany_c$ is an indicator for Germany, and $Post\ NetzDG_t$ is an indicator for the quarters after the NetzDG. All regressions include platform (γ_i), country (ω_c), and quarter (δ_t) fixed effects. The main coefficient of interest β_1 measures relative changes in the usage of moderated platforms vis-à-vis unmoderated platforms in Germany relative to changes of usage of the same platforms in other countries.

The identifying assumption underlying these regressions is that without the NetzDG, the relative use of moderated and unmoderated platforms in Germany would have followed similar trends as in other countries (Olden and Møen, 2022). In other words, we do not require that the moderated and unmoderated platforms follow similar usage trends, only that the relative usage of the platforms across countries would remain stable in the absence of the NetzDG. Also note that the triple-difference design allows us to abstract from differences in the usage of platforms in different countries and any usage changes within countries. Moreover, this design is robust to spillovers between countries or within platforms.

We provide support for the parallel trends assumption by testing for pre-trends in Figure 9. We find that relative platform usage in Germany followed similar trends, before the NetzDG, when compared to the other countries in our sample. Note that this figure begins in 2017q1 as these are the earliest quarters for which Semrush web traffic data exist. After the passage of the NetzDG from 2017q4 onwards, we find overall significantly positive estimates for the usage of the moderated platforms in Germany.

Figure 9: The Effect of the NetzDG on Platform Usage



Notes: This figure plots coefficients from event study versions of Equation (2). The dependent variable is the log number of users. The omitted category is the 3rd quarter of 2017, the quarter of the enactment of the NetzDG (indicated with the vertical line). The whiskers indicate 95% confidence intervals based on standard errors clustered by country.

These findings are confirmed by the regression estimates in Table 2. The estimates suggest that the quarterly usage of moderated platforms increased by 8% when measured by the number of users (columns (1) and (2)) and by 9% based on the total number of visits (columns (3) and (4)). The estimates remain unchanged when we include additional interacted fixed effects (columns (2) and (4)).

Table 2: Regression Estimates: NetzDG and Platform Usage

	Dep. var.: Log(Users)		*Dep. var.: Log(Visits)*	
	(1)	(2)	(3)	(4)
Germany × Platform × Post	0.081***	0.081***	0.089***	0.089***
	(0.022)	(0.022)	(0.029)	(0.029)
Germany × Post	0.091**		0.066	
	(0.039)		(0.040)	
Country FE	Yes		Yes	
Year-Quarter FE	Yes		Yes	
Platform FE	Yes		Yes	
Country × Year-Quarter FE		Yes		Yes
Platform × Year-Quarter FE		Yes		Yes
Country × Platform FE		Yes		Yes
Observations	3,024	3,024	3,024	3,024
Pre-Period Mean of DV	15.86	15.86	17.62	17.62
R^2	0.93	0.99	0.92	0.99

Notes: This table presents the results of estimating Equation (2), where the dependent variable is the log number of unique users or the log of the total visits to a website. *Germany* is an indicator variable equal to 1 for Germany, and 0 otherwise. $Platform$ is an indicator equal to 1 for the platforms targeted by the NetzDG (Instagram, Twitter, YouTube, and Facebook), and 0 for those that were not (Amazon, Netflix, Wikipedia). *Post* is a dummy variable equal to 1 for observations after 2017q3, the quarter of the NetzDG enactment. Standard errors in parentheses are clustered at the country level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Why did the NetzDG result in an increase in traffic to these websites? The persistent effects in Figure 9 suggest that the pattern is unlikely to be explained by a temporary increase in the salience or popularity of the treated platforms due to the passage of the law. Instead, these findings are consistent with an increase in demand for these websites, suggesting that most users prefer a marginal increase in moderation. These results are consistent with the evidence in Jiménez Durán (2022), who documents a positive effect of reporting hate speech on the engagement of users who are attacked by hateful posts. Appendix Figure C.4 provides further estimates at the platform level. Platforms with a higher reliance in user reports for content moderation (Instagram and Twitter) are also the ones where the effect is stronger.[24]

[24]As argued by Jiménez Durán (2022), this pattern is to be expected, given that platforms moderate up to a point where the marginal benefit (an increase in user engagement) equals marginal cost. We expect platforms with a higher reliance on user reports to have a higher marginal cost be-

Lastly, Appendix Figure C.5 displays a specification curve, where we compute the sensitivity of our main estimates (those of column (2) in Table 2) to all possible combinations of treated and control platforms. As the figure shows, the estimates remain positive and significant through most platform combinations, and the baseline effect size we estimate corresponds roughly to the median effect obtained across all specifications. This exercise suggests that it is possible to rule out a negative effect of the NetzDG on platform usage.

5 Offline Effects of the NetzDG

The second part of the paper investigates the offline effects of the NetzDG. The analysis again proceeds in three steps. First, we study if the NetzDG-induced decrease in online toxicity also led to a reduction in the prevalence of anti-refugee hate crimes. Second, we investigate two potential mechanisms that could explain changes in the anti-refugee incidents. Lastly, we analyze the impact of the NetzDG on overall hate crimes in Germany using a cross-country synthetic control design.

5.1 Did the NetzDG Reduce Anti-Refugee Hate Crimes?

To estimate the effect of the NetzDG on anti-refugee hate crimes, we exploit variation in the exposure of different German municipalities to anti-refugee content. Intuitively, we expect places with a higher exposure to this type of content to be disproportionately affected by the NetzDG relative to places with a lower exposure.

Empirical Strategy

This intuition gives rise to the following empirical strategy:

$$y_{it} = \theta \cdot AfD\ Users\ p.c._i \times Post\ NetzDG_t + \mathbf{X_{it}'}\beta + \gamma_i + \delta_t + \epsilon_{it}, \tag{3}$$

cause they likely rely more on human reviewers than platforms that proactively remove content (which typically rely heavily on automated systems). See, for example, Meta's proactive detection strategy: `https://transparency.fb.com/policies/improving/proactive-rate-metric/`. In 2019q3 (the earliest for which there is data), 55.9% of content violating Instagram's rules was found through user reports. This compares to 19.3% for Facebook (`https://transparency.fb.com/reports/community-standards-enforcement/hate-speech`). On YouTube, non-automated video removals amounted to 20.4% of removals (but this figure also includes offenses other than hate speech, see `https://transparencyreport.google.com/youtube-policy/removals`). In contrast, in 2020, close to half of violating content on Twitter was flagged by humans (`https://www.fastcompany.com/90528941/twitter-automatically-flags-more-than-half-of-all-tweets-that-violate-its-rules`).

where our main outcome of interest, y_{it}, is the inverse hyperbolic sine of the number of anti-refugee incidents in municipality i in quarter t.[25] The main independent variable is the interaction between the number of AfD Facebook users per capita ($AfD\ Users\ p.c._i$) and a time dummy ($Post\ NetzDG_t$) which is equal to one for the period after 2017q3, when the NetzDG was enacted. The regression includes a full set of municipality and time fixed effects. The municipality fixed effects control for any baseline difference in the number of anti-refugee incidents (e.g., due to the higher presence of refugees), while the time fixed effects account for any Germany-wide change in the number of anti-refugee incidents (e.g., due to national news events).

Table A.5 plots the mean and standard deviation of a large number of municipality characteristics, binned by quartiles of our exposure variable, $AfD\ Users\ p.c._i$. More exposed municipalities tend to be somewhat larger and more likely to vote for the AfD, Linke, or Green party, but these differences are quantitatively small. To control for potential other drivers of trends in hate crimes over time, the vector ($\mathbf{X_{it}}$) includes control variables, which we also interact with the $Post\ NetzDG_t$ dummy. We cluster standard errors at the county level.[26]

As is standard for difference-in-differences designs, our identifying assumption is that in the absence of the NetzDG, municipalities with different prior exposures to hate speech on social media would have experienced a similar trend in hate crimes. The coefficient θ, therefore, measures the extent to which the NetzDG was associated with a differential change in the number of anti-refugee incidents in municipalities with a higher exposure to anti-refugee content on Facebook.

Results

Table 3 shows our main results. Column (1) contains estimates of our baseline specification using Equation (3), controlling only for log population interacted with the *Post* indicator to control for changes in hate crimes due to population differences. In the remaining columns, we add controls for potential confounders. Across these different specifications, the point estimates remain stable and indicate that municipalities with a one standard deviation higher number of AfD Facebook users per capita saw a 1% larger reduction in hate crimes. As a benchmark, Müller and Schwarz (2021) find that a one standard deviation increase in AfD Facebook users per capita is associated with a 10% higher probability of a weekly anti-refugee incident relative to the mean. Our estimate on the effect of the NetzDG seems plausible given the 15% reduction in hateful online content we identified for AfD users in the previous section.

[25] In Appendix Table D.2, we show that the results are robust to other variable transformations.
[26] In Appendix Table D.3, we show robustness for alternative levels of clustering.

In column (2), we control for the vote share of the AfD and all other major parties at the municipality level. These account for changes in anti-refugee incidents around the time of the NetzDG that can be explained by the political leaning of a municipality. We find that the coefficient for the AfD vote share is positive and significant. This result highlights the clear distinction between offline support for the AfD and online exposure to hateful content, the former of which is unaffected by the NetzDG. Controlling for the AfD vote share further allows us to mitigate concerns about many other contemporaneous shocks. The reason is that shocks other than the NetzDG that may disproportionately reduce anti-refugee attacks in right-leaning areas should affect AfD voters similarly to AfD Facebook users. Our results instead point toward the importance of an online channel.

Column (3) adds Facebook users per capita in a municipality as a control to account for changes in anti-refugee incidents that could be explained by unobservable confounders that correlate with a municipality's affinity to social media. In a similar spirit, we add a control for the access to broadband internet in column (4), which is our preferred specification, since it controls for population, voting, and (social) media consumption. The coefficients on Facebook users per capita and broadband internet access are small and statistically indistinguishable from 0. In other words, after accounting for the exposure to far-right Facebook usage, a town's social media or internet penetration does not matter for its elasticity with respect to the NetzDG. This finding suggests that, in line with our hypothesis and the evidence in the first part of the paper, the NetzDG mattered for people who were exposed to anti-refugee content instead of the effects being driven by access to social media or the internet. Finally, in column (5), we include a wealth of additional control variables (see Appendix A for details), all of which we interact with the *Post* indicator. The inclusion of these 19 additional control variables again has little impact on the magnitude, sign, and statistical significance of our main estimate.

Event Study

Figure 10 visualizes the coefficients from an event study version of regression Equation (3), with 2017q3 (the quarter when the NetzDG was enacted) as the excluded period. We find no evidence for pre-existing trends in this specification. The pre-period coefficients are statistically insignificant and close to 0. We only observe a statistically significant reduction in the number of anti-refugee incidents after the increase of content moderation efforts in 2017q4. Moreover, this negative effect appears to be persistent and stable over the two years following the NetzDG.

Table 3: Effect of NetzDG on Anti-Refugee Hate Crime

	Dep. var.: Asinh(Anti-Refugee Hate Crimes)				
	(1)	(2)	(3)	(4)	(5)
AfD Facebook users p.c. (std) × Post	-0.012***	-0.009***	-0.009***	-0.009***	-0.008***
	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)
AfD vote share (std) × Post		0.034***	0.034***	0.036***	0.031***
		(0.012)	(0.012)	(0.012)	(0.012)
Facebook users p.c (std) × Post			0.002	0.002	0.003
			(0.002)	(0.002)	(0.002)
Broadband internet (std) × Post				0.005	0.001
				(0.003)	(0.004)
Municipality FE	Yes	Yes	Yes	Yes	Yes
Year-Quarter FE	Yes	Yes	Yes	Yes	Yes
Ln(Pop.) × Post	Yes	Yes	Yes	Yes	Yes
Election Controls × Post		Yes	Yes	Yes	Yes
All Controls (19) × Post					Yes
Observations	71,456	71,008	71,008	71,008	68,736
Pre-Period Mean of DV	0.12	0.12	0.12	0.12	0.12
R^2	0.44	0.44	0.44	0.44	0.45

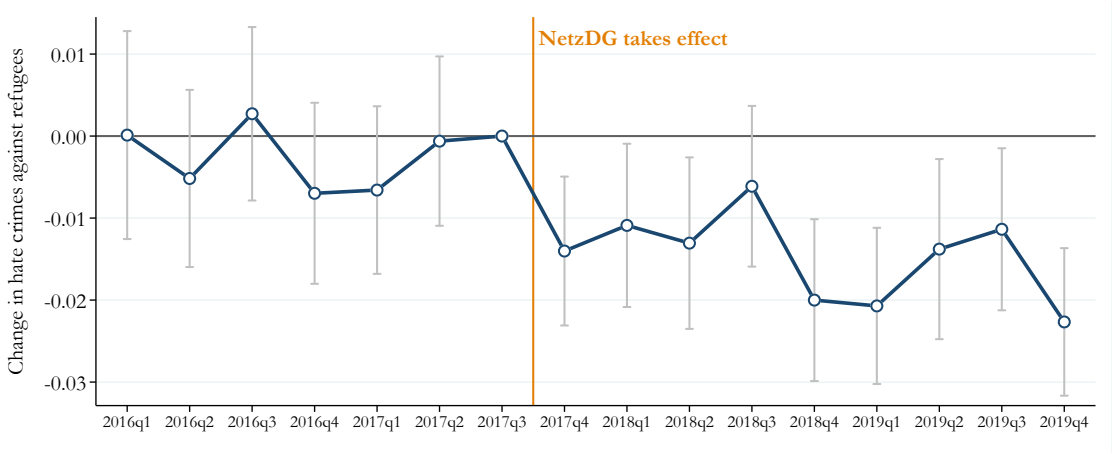
Notes: This table presents the results of estimating Equation (3), where the dependent variable is the inverse hyperbolic sine of the number of anti-refugee hate crimes in a municipality in a given quarter. *AfD Facebook users p.c. (std)* is the number of AfD Facebook followers per capita, standardized to have a mean of 0 and a standard deviation of 1 to ease interpretation. All regressions include municipality and quarter fixed effects as well as a control for the natural logarithm of population, interacted with *Post*. See text for a detailed description of the additional control variables. Robust standard errors in parentheses are clustered by county. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Alternative Explanations

As with any difference-in-differences estimation, identification requires the absence of other contemporaneous shock that differentially affects areas depending on the presence of AfD Facebook users. As discussed in the Section 4, two possible candidates for such shocks could be (1) the end of the refugee crisis, and (2) the 2017 federal election. We have already shown that these events cannot explain the online patterns above. In the following, we discuss why they are also unlikely to drive our findings on the offline consequences of content moderation.

First, our findings cannot be easily explained by some form of mean reversion in the number of anti-refugee incidents due to the end of the refugee crisis in Germany. As discussed in Section 2, the inflow of refugees to Germany had already stopped in March 2016 when the EU struck a deal with Turkey to prevent the further entry of refugees from Syria to Europe. Therefore, the total number of refugees was nearly constant around the introduction of the NetzDG. Further, the effect we find occurs over a year after this important demarcation point of the refugee crisis. It is further worth noting that the exposure measure we use is essentially

Figure 10: Event Study Hate Crime



Notes: This figure plots the coefficients from running an event study version of regression Equation (3). The dependent variable is the inverse hyperbolic sine of the number of anti-refugee incidents. The omitted category is the 3rd quarter of 2017, the quarter before the passing of the NetzDG (indicated with the vertical line). The whiskers indicate 95% confidence intervals based on standard errors clustered by county.

uncorrelated with the number of refugees (see Appendix Table A.5). As a result, including the number of refugees in a municipality interacted with the $Post\ NetzDG_t$ indicator as a control does not change the estimates (see column 5 Table 3). Moreover, any such mean reversion should also affect municipalities with many AfD voters in a similar way, which is rejected by the estimated positive coefficient on the AfD vote share.

Second, the 2017 federal election is unlikely to drive our findings. When we include controls for the electoral results of all major German parties in our regressions, the magnitude and statistical significance of our coefficient of interest barely changes. Moreover, the positive coefficient for the AfD vote share contradicts the idea that the end of the election period was associated with a drop in the number of anti-refugee incidents. Instead, these results would be consistent with the interpretation that the unexpectedly strong showing of the AfD in the 2017 federal elections, where it became the third strongest party and the first far-right party in the German parliament since 1945, may have emboldened its supporters. This hypothesis meshes well with evidence in Bursztyn et al. (2020), who document that the election of Donald Trump in the United States increased people's willingness to publicly express xenophobic views.[27]

[27]Note that this argument does not rely on assumptions that AfD Facebook users and AfD voters are identical in their ideological convictions, even though there is a positive correlation with the presence of both with anti-refugee incidents in the pre-period. We only require that the effect works in the same direction for both groups. For example, any hypothetical moderation in rhetoric by AfD politicians is unlikely to reduce refugee incidents for Facebook users while increasing them for AfD voters.

Finally, our results effectively exploit cross-sectional exposure in the residual variation in AfD Facebook usage that is not explained by either AfD or Facebook affinity. This strategy makes it unlikely that any other event that may have occurred contemporaneously with the passage of the NetzDG biases our estimates. In order for such an event to be a potential confounder, it would have to simultaneously reduce anti-refugee incidents in municipalities with many AfD Facebook users, increase anti-refugee incidents in municipalities with AfD voters, and leave municipalities with many Facebook users unaffected. It is also worth noting that there is no evidence of changes in the reporting of anti-refugee incidents around the time of the NetzDG.

Robustness

To further probe the robustness of our findings, we perform additional checks. First, Online Appendix Table D.1 shows that, with the exception of cases of arson (which are rare), the NetzDG affected all categories of anti-refugee incidents (i.e., assault, demonstration, suspected attacks, and other miscellaneous property attacks). The strongest response is for assaults and other property attacks. The effect on severe incidents such as assaults, which seem difficult to "fake," also makes it less likely that the estimates capture differential changes in the likelihood an incident gets reported. Besides, any overall change in the reporting of incidents would be absorbed by the time fixed effects. It is also worth noting that the passage of the NetzDG and the surrounding debate on hate crime should, if anything, make it more likely for victims to report incidents.

Table 4 presents a battery of additional robustness exercises. Column (2) shows robustness to the inclusion of federal state $\times$ quarter fixed effects (see column (2)). This specification exploits variation within the same federal state at the same point in time, and hence accounts for any potential changes in law enforcement that might have been introduced by the state governments. These fixed effects will also absorb any differential shock that might affect a specific federal state (e.g., local elections). Column (3) excludes January and February 2016 from the data, which constitute the largest spike in anti-refugee incidents. This exclusion leaves the estimates unchanged and highlights that the findings are not driven by outliers. Similarly, the findings are robust to excluding municipalities without anti-refugee incidents, without AfD users, or with few refugees per capita (columns (4), (5), and (6), respectively). Throughout these exercises, the estimates remain statistically significant, making it unlikely that they are driven by zeroes in the dependent or independent variables.

Third, Appendix Table D.2 shows that the estimates are robust independently of the functional form of the dependent and independent variables. In particular, we explore transformations of the dependent variable (refugee attacks) in inverse hyperbolic sine (baseline),

Table 4: Robustness Tests

	Dep. var.: Asinh(Anti-Refugee Hate Crimes)					
	Baseline (1)	Federal State × Quarter FE (2)	Exclude Q1 2016 (3)	Exclude Attack= 0 (4)	Exclude AfD User= 0 (5)	Exclude Few Refugees (6)
AfD Facebook users p.c. (std) × Post	-0.009***	-0.008***	-0.009***	-0.016***	-0.009***	-0.016***
	(0.002)	(0.002)	(0.002)	(0.005)	(0.003)	(0.004)
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes
Year-Quarter FE	Yes	Yes	Yes	Yes	Yes	Yes
Fed. State × Quarter FE		Yes				
Ln(Pop/) × Post	Yes	Yes	Yes	Yes	Yes	Yes
AfD vote share × Post	Yes	Yes	Yes	Yes	Yes	Yes
Election Controls × Post	Yes	Yes	Yes	Yes	Yes	Yes
Facebook users p.c × Post	Yes	Yes	Yes	Yes	Yes	Yes
Broadband internet × Post	Yes	Yes	Yes	Yes	Yes	Yes
Observations	71,008	71,008	66,570	36,384	64,736	56,656
Pre-Period Mean of DV	0.12	0.12	0.10	0.23	0.12	0.14
R^2	0.44	0.45	0.45	0.42	0.44	0.46

Notes: This table presents the results of estimating municipality-quarter-level regressions as in Equation (3), where the dependent variable is the inverse hyperbolic sine of the number of anti-refugee hate crimes. *AfD Facebook users p.c. (std)* is the number of AfD Facebook followers per capita, standardized to have a mean of 0 and a standard deviation of 1 to ease interpretation. All regressions include municipality and quarter fixed effects, as well as controls for the logarithm of population, the AfD vote share, Facebook users per capita, and broadband internet access, all interacted with *Post*. Robust standard errors in parentheses are clustered by county. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

counts, or the log number of refugee incidents per capita. Neither of these changes alter our findings (see columns (1-3)). Columns (4-6) then replace the main independent variable with an indicator of whether a municipality has an above-median number of AfD users per capita. This exercise serves three purposes. First, it allows us to rule out concerns about outliers in the number of AfD users per capita. Second, this dummy specification does not rely on functional form assumptions, because it simply picks up changes in the mean number of anti-refugee incidents after the NetzDG in a canonical difference-in-differences setting. Third, this transformation also alleviates concerns that the findings are driven by heterogeneous treatment effects in the two-way fixed effects estimation (De Chaisemartin and d'Haultfoeuille, 2023), as our results also hold in this dummy specification.

Fourth, we repeat the analysis based on the number of followers the AfD has in a municipality on Twitter rather than Facebook. Appendix Table D.4 shows that the results are highly similar with this alternative measure of exposure to the NetzDG. Appendix Figure D.2 also presents the corresponding event study estimates.

Fifth, we perform a leave-one-out analysis excluding one municipality at a time. The results are shown in Appendix Figure D.1. The estimates are highly stable throughout. As such, our findings do not appear to be driven by any particular municipality.

Finally, Appendix Table D.3 shows that the estimates remain statistically significant irrespective of the level of clustering of the standard errors. Specifically, the results are similar

when standard errors are clustered at (1) the county level (baseline), (2) the county and quarter level, (3) the municipality level, or (4) the municipality and quarter level.

The Role of User Engagement

We also investigate the role of user interactions. In principle, the NetzDG could affect the prevalence of anti-refugee incidents by changing the willingness of either the consumers or producers of anti-refugee online content to commit acts of violence against refugees. While we do not have clear-cut measures of the production and consumption of online hate, we can investigate heterogeneity in our estimates depending on the extent of user engagement with the AfD's Facebook page.

Our idea is the following. The number of right-wing social media users is a plausible proxy for exposure or consumption of such content in a municipality. If we additionally observe that the intensity of user interactions matters over and above mere exposure, this may be indicative of a role for the production of hateful online content. In our setting, we create proxies for user engagement (or "production" of online hate) by the average number of posts, comments, likes, and shares by each AfD users in a municipality, defined before the passing of the NetzDG.

Table 5 plots the results of regressions where we include our different measures of user engagement. Note that these regressions are only estimated for municipalities for which we can identify at least one AfD user, because these measures are not defined in the case of zero users. The results suggest that municipalities with stronger right-wing social media interactions saw a larger decline in anti-refugee hate crimes following the NetzDG, even after controlling for the exposure (or "consumption") of such content as measured by the number of AfD users on Facebook. The coefficient in column (1) suggests that a one standard deviation increase in the number of posts per AfD user is associated with an additional 0.5 percentage point reduction in the number of anti-refugee hate crimes.

Although the interactions with our proxies of user engagement are always statistically significant, we also find that the number of AfD Facebook users per capita in the first row (our proxy for the "consumption" of online hate) still has predictive power. This pattern holds for all measures of engagement, such as likes, comments, or shares. Taken together, these results suggest that the NetzDG affected the likelihood of committing anti-refugee acts, especially for places with considerable engagement on right-wing social media, perhaps reflecting a role for both consumers and producers of anti-refugee content.

Table 5: Heterogeneity by User Activity

	Dep. var.: Asinh(Anti-Refugee Hate Crimes)			
	(1)	(2)	(3)	(4)
AfD Facebook users p.c. (std) × Post	-0.022***	-0.022***	-0.022***	-0.022***
	(0.004)	(0.004)	(0.004)	(0.004)
Post per AfD User (std) × Post	-0.005***			
	(0.001)			
Likes per AfD User (std) × Post		-0.005***		
		(0.001)		
Comments per AfD User (std) × Post			-0.004***	
			(0.001)	
Shares per AfD User (std) × Post				-0.004***
				(0.002)
Municipality FE	Yes	Yes	Yes	Yes
Year-Quarter FE	Yes	Yes	Yes	Yes
Ln(Pop.) × Post	Yes	Yes	Yes	Yes
Observations	57,008	57,008	57,008	57,008
Pre-Period Mean of DV	0.14	0.14	0.14	0.14
R^2	0.45	0.45	0.45	0.45

Notes: This table presents the results from estimating Equation (3) for municipalities with at least one AfD Facebook user. The dependent variable is the inverse hyperbolic sine of the number of anti-refugee hate crimes in a municipality and quarter. *AfD Facebook users p.c. (std)* is the number of AfD Facebook followers, standardized to have a mean of 0 and a standard deviation of 1 to ease interpretation. We additionally include different measures of Facebook activity per AfD user before the NetzDG in regressions, also standardized to have a mean of 0 and a standard deviation of 1. All regressions include municipality and quarter fixed effects, as well as a control for the logarithm of population interacted with *Post*. Robust standard errors in parentheses are clustered by county. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

5.2 Potential Mechanisms

We now shed light on two different mechanisms that may partially explain why an increase in content moderation could reduce hate crimes. First, we analyze whether the NetzDG made collective action against refugees more difficult. Second, we study whether the NetzDG influenced attitudes toward refugees.

Did the NetzDG Affect Collective Action?

Following the literature on the effects of social media on collective action (e.g., Enikolopov et al., 2020; Manacorda and Tesei, 2020; Fergusson and Molina, 2021), we investigate whether the NetzDG was able to interrupt the ability of potential perpetrators to coordinate anti-refugee incidents. As an example, the NetzDG could make it harder to learn about the willingness of others to carry out acts of violence against refugees. Recall from the result in column (6) of Table C.7 that the NetzDG made the tone of tweets less threatening. To examine a potential coordination mechanism, we rerun our main analysis but split anti-refugee incidents based on the number of perpetrators. Note that we could only hand-code the number

of perpetrators for 9% of the anti-refugee incidents in our data and thus reduces the mean of the dependent variable, which mechanically leads to smaller coefficients in these regressions.

Table 6 presents the results from this analysis. Panel (a) shows the results for anti-refugee incidents with a single perpetrator, whereas Panel (b) shows the estimates for multiple perpetrators. While the estimates in both panels are statistically significant, the effect of the NetzDG on incidents with multiple perpetrators is, in all cases, twice as large as for incidents with a single perpetrator. These findings highlight the "social" component of anti-refugee incidents and suggest social media may help facilitate collecting action (in this case, violent attacks on refugees). This evidence is also in line with previous work by Müller and Schwarz (2021), who document a stronger effect of social media on hate crimes with multiple perpetrators, as well as the evidence in Bursztyn et al. (2019) of a coordination mechanism of social networks in the Russian context.

Following the evidence from Enikolopov et al. (2020), we investigate if collective action effects are stronger in larger and more urban municipalities. The idea underlying this test is that the value of social media for collective action increases with city size due to weaker social ties offline. The results are reported in Appendix Table D.5. We consistently find that the effects are driven by municipalities with an above-median population or above-median population density. The effect of the NetzDG is again stronger for incidents with multiple perpetrators.

We additionally investigate if the effect of the NetzDG on collective action is driven by information provision. To this extent, we search the refugee tweets in our data for mentions of locations in Germany. In line with the findings of Müller and Schwarz (2021), we find no evidence for tweets that actively coordinate incidents against refugees or mention refugee locations. These results suggest that the effect of the NetzDG on anti-refugee incidents is more likely the result of reduced emotional contagion rather than information provision.

Did the NetzDG Affect Attitudes Towards Refugees?

The NetzDG may also have decreased hate crimes because it changed attitudes toward refugees, for example, by reducing the animus of social media users towards refugees. To examine this idea, we use data from the German Socio-Economic Panel (GSOEP) (Goebel et al., 2019). Specifically, we exploit the panel nature of GSOEP to study within-person changes in attitudes towards refugees using a regression of the following form:

$$y_{it} = \theta \cdot Social\ Media\ User \times Post\ NetzDG_t + \mathbf{X'_{it}}\beta + \gamma_i + \delta_t + \epsilon_{it}, \tag{4}$$

Table 6: The NetzDG and Hate Crime Coordination

	Dep. var.: Asinh(Anti-Refugee Hate Crimes)				
	(1)	(2)	(3)	(4)	(5)
Panel (a): Single Perpetrators					
AfD Facebook users p.c. (std) × Post	-0.002***	-0.001**	-0.001**	-0.001**	-0.001**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Observations	71,456	71,008	71,008	71,008	68,736
Pre-Period Mean of DV	0.005	0.005	0.005	0.005	0.005
R^2	0.14	0.14	0.14	0.14	0.14
Panel (b): Multiple Perpetrators					
AfD Facebook users p.c. (std) × Post	-0.004***	-0.002**	-0.002**	-0.002**	-0.002**
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Municipality FE	Yes	Yes	Yes	Yes	Yes
Year-Quarter FE	Yes	Yes	Yes	Yes	Yes
Ln(Pop/) × Post	Yes	Yes	Yes	Yes	Yes
AfD vote share × Post		Yes	Yes	Yes	Yes
Election Controls × Post		Yes	Yes	Yes	Yes
Facebook users p.c × Post			Yes	Yes	Yes
Broadband internet × Post				Yes	Yes
All Controls (19) × Post					Yes
Observations	71,456	71,008	71,008	71,008	68,736
Pre-Period Mean of DV	0.012	0.012	0.012	0.012	0.012
R^2	0.20	0.20	0.20	0.20	0.21

Notes: This table presents the results of estimating Equation (3), where the dependent variable is the inverse hyperbolic sine of the number of anti-refugee hate crimes in a municipality in a given quarter. We restrict the sample to hate crimes committed by single perpetrators in panel (a) and multiple perpetrators in panel (b). *AfD Facebook users p.c. (std)* is the number of AfD Facebook followers per capita, standardized to have a mean of 0 and a standard deviation of 1 to ease interpretation. All regressions include municipality and quarter fixed effects as well as a control for the natural logarithm of population, interacted with *Post*. See text for a detailed description of the additional control variables. Robust standard errors in parentheses are clustered by county. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

where y_{it} is the response to one of five questions on the impacts of refugees on the economy and society. As described above, we recode these questions into indicator variables such that 1 represents a positive attitude towards refugees. We additionally create an index capturing the average response of a respondent across these five questions. $Social\ Media\ User$ is an indicator for respondents who use social media at least once a week. $Post\ NetzDG_t$ is a dummy for the period after the NetzDG. As the data only contains two survey waves, $Post\ NetzDG_t$ is 1 for the year 2018 and 0 for 2016. $\mathbf{X_{it}}$ contains controls for the gender and age of the respondents, which we interact with the $Post$ indicator. Finally, γ_i and δ_t are full sets of respondent and survey wave fixed effects. As a result, β measures whether respondents who use social media developed more positive attitudes towards refugees between 2016 and 2018 relative to respondents who did not use social media.

Table 7: Changes in Attitudes Towards Refugees

	Dep. var.: Refugees are ...					
		Positive for the			A Chance in the	
	Index	Economy	Culture	Place of Living	Short-term	Long-term
	(1)	(2)	(3)	(4)	(5)	(6)
Panel (a): All Respondents						
Social Media User × Post	-0.008	0.001	-0.011	0.001	-0.019**	-0.010
	(0.006)	(0.009)	(0.009)	(0.008)	(0.009)	(0.009)
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes
Survey Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	36,698	36,296	36,280	36,254	36,268	36,144
Pre-Period Mean of DV	0.50	0.60	0.57	0.53	0.24	0.54
R^2	0.82	0.73	0.75	0.75	0.66	0.74
Panel (b): AfD Voters						
Social Media User × Post	-0.002	0.029	-0.034	-0.009	0.004	0.004
	(0.017)	(0.034)	(0.030)	(0.024)	(0.019)	(0.026)
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes
Survey Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2,566	2,550	2,558	2,548	2,554	2,550
Pre-Period Mean of DV	0.16	0.25	0.18	0.15	0.06	0.16
R^2	0.71	0.64	0.64	0.66	0.55	0.66

Notes: This table presents the results of estimating Equation (4), where the dependent variables are different measures for positive attitudes towards refugees. *Social Media Users* is an indicator for respondents who use social media at least once a week. All regressions include individual and survey year fixed effects as well as controls for the respondent's gender and age, interacted with *Post*. See the text for a detailed description of the variables. Robust standard errors in parentheses are clustered by individual. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 7 plots the results. Overall, there is no evidence for positive changes in attitudes for the period after the NetzDG. All estimates are small and, except one, statistically insignificant. The only significant estimate is *negative*, which would reject the hypothesis that the NetzDG improved attitudes towards refugees. These null results hold both for all GSOEP respondents (Panel a) as well as respondents who express support for the AfD (Panel b). As an additional test, we also investigate changes in pro-refugee *actions* (as opposed to opinions) in Appendix Table D.6. The estimates are again mostly small and, if anything, negative. Taken together, these findings provide evidence against the idea that the NetzDG has caused a reduction in anti-refugee incidents primarily by changing attitudes towards refugees.

5.3 Synthetic Control Estimates

As the last piece of our analysis, we provide additional evidence for the offline effects of the NetzDG based on synthetic control estimates. This serves two purposes. First, it lends further credence to our main findings using a completely different data source and empirical

strategy. Second, the synthetic control estimates allow us to investigate the effect on the total (non-refugee related) number of hate crimes in Germany, which are only available at the country-year level. More specifically, we build a synthetic control group for Germany using data from 21 donor countries from the OSCE, following the methodology of Abadie and Gardeazabal (2003) and Abadie et al. (2010). The dependent variable is the yearly number of hate crimes per 10,000 inhabitants, and we use as predictors the full path of lagged outcomes, as recommended by Ferman et al. (2020). Because some of the donor countries changed their data collection in the pre-period, we add as a predictor an indicator of whether there was a change in measurement. Because the NetzDG became law in the fourth quarter of 2017, we define 2017 as the treatment year. This approach is more conservative than using 2018 as the treatment year since backdating the intervention does not mechanically bias the estimator (Abadie, 2021).

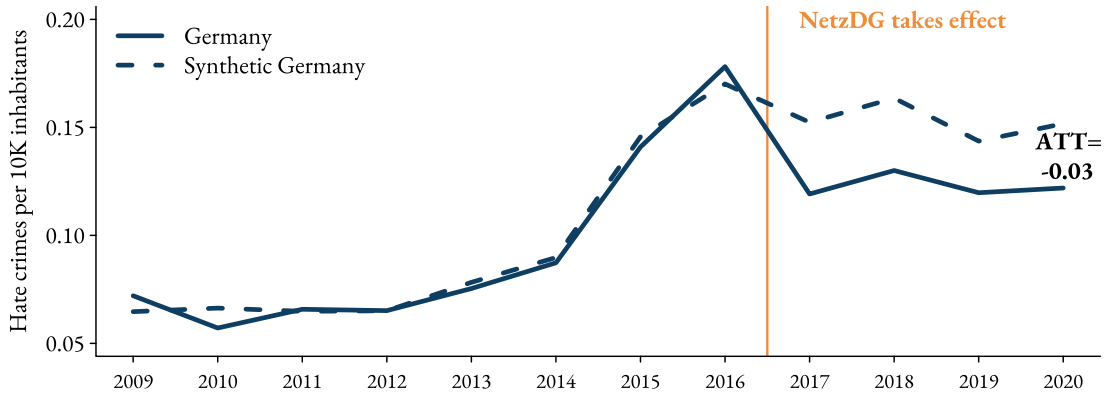
Figure 11 reports the main estimates from this exercise. This figure shows that the number of hate crimes per 10,000 inhabitants in the synthetic Germany based on the 21 donor countries by construction closely tracks the observed hate crimes until the year the NetzDG was enacted. After the NetzDG goes into force, we find a drop in the number of hate crimes relative to the synthetic control. The average treatment effect (ATE) in the 2018-2020 post-period is -0.0301 hate crimes per 10,000 inhabitants, or 250 fewer hate crimes per year. Appendix D.3. presents additional information, such as the weights used to construct the synthetic control (Table D.7) and the pre-period balance of the predictors (Table D.8).[28]

We can reject the null hypothesis that the ATT is non-negative with a p-value of 0.045, constructed based on in-space placebo tests as in Abadie et al. (2010), where "placebo effects" are computed assuming that each of the donor countries is treated. Intuitively, this exercise shows that the magnitude of the treatment effect in Germany is an outlier relative to the placebo effects estimated among countries in the donor pool. Figure D.5 provides visual evidence of this intuition by plotting the histograms of the (one-sided) ratio of the mean square predicted errors (MSPE) after vs. before the NetzDG in Germany and in the donor countries.

Table D.9 shows that this result is robust to a battery of additional checks. First, we investigate alternative ways of dealing with missing data (no interpolation or including a dummy for interpolated values). Second, we explore alternative transformations of the outcome variable (logarithm and levels). Third, we consider alternative end dates, which result in a different donor pool based on differences in data availability. Fourth, we consider alternative sets of donor countries (leaving out donor countries or restricting our estimates to

[28]The weights overall seem intuitive, with countries like Poland, Italy, and Austria receiving a large weight (close to 10% each). The one outlier is the large weight of Lithuania (55%). In robustness checks in Table D.9, we confirm that our results do not change when we remove Lithuania from the donor pool.

Figure 11: Evolution of Hate Crimes in Germany vs. Synthetic Germany



Notes: This figure presents the evolution of hate crimes per 10,000 inhabitants in Germany and a synthetic Germany. The synthetic control uses all lagged outcomes as predictors, as well as a dummy variable indicating whether there were measurement changes in the pre-period.

OECD members). Overall, the results remain similar throughout these robustness checks, suggesting that the NetzDG contributed to reducing the aggregate number of hate crimes in Germany relative to other countries.

As a placebo exercise, we replicate the estimation on the number of homicides per 10,000 inhabitants, based on the assumption that it is unlikely that the NetzDG impacted the overall homicide rate. If our results were driven by other policies implemented by Germany that coincided with the NetzDG and also impacted hate crimes, such as a change in law enforcement, we would also expect to see an effect on this outcome variable. As Figure D.6 shows, there is no evidence of an effect on homicides; the estimate is positive in some years and negative in others. The effect size is only one-fourth of the estimate for hate crimes (relative to the level of pre-treatment outcomes). Moreover, as opposed to our estimates on hate crimes, the magnitude of the effect is small compared to the placebo effect on the donor countries, which is reflected in a p-value of 0.44.

6 Discussion

Much attention has been devoted to the spread of hateful content on social media. The controversial German NetzDG was in large part a reaction to the prevalence of hateful messages on social media platforms and the perceived limited effort of these platforms to moderate this content. By leveraging this unique quasi-experiment, this paper is the first to show that content moderation, induced by regulation, can indeed achieve its primary aim

of reducing hateful sentiments online and decreasing the incidence of hate crimes against minorities offline.

While reducing hate is undoubtedly an important aim, we want to caution against taking this finding as blanket support for content moderation. This study does not and cannot evaluate the full schedule of costs and benefits of online censorship and its potential impact on legitimate online debate. For example, one of the main reasons why the NetzDG has been controversial is its potential misuse to undermine freedom of expression and stifle political dissent (Kaye, 2018). We do not find evidence that the law increased online discussions of censorship or that users disengaged from controversial political issues. However, Figure A.3 shows that an index measuring freedom of expression in Germany based on expert opinions decreased after the passage of the law, moving it from third place in 2016 (between Switzerland and Belgium) to seventeenth place in 2020 (between New Zealand and Uruguay). While it is unclear whether this decrease is driven by the NetzDG, it highlights the need for more research to understand the effects of content moderation policies on freedom of expression and offline political discussion. As such, we believe our findings should best be interpreted as a starting point for understanding the online and offline effects of content moderation on social media.

References

Abadie, A. (2021). Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects. *Journal of Economic Literature 59*(2), 391–425.

Abadie, A., A. Diamond, and J. Hainmueller (2010). Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association 105*(490), 493–505.

Abadie, A. and J. Gardeazabal (2003). The Economic Costs of Conflict: A Case Study of the Basque Country. *American Economic Review 93*(1), 113–132.

Acemoglu, D., T. A. Hassan, and A. Tahoun (2017, 08). The Power of the Street: Evidence from Egypt's Arab Spring. *The Review of Financial Studies 31*(1), 1–42.

Adena, M., R. Enikolopov, M. Petrova, V. Santarosa, and E. Zhuravskaya (2015). Radio and the Rise of The Nazis in Prewar Germany. *The Quarterly Journal of Economics 130*(4), 1885–1939.

Alba, D. and K. Wagner (2023, January). Twitter cuts more staff overseeing global content moderation. *Bloomberg News*. Accessed: 2025-02-14.

Allcott, H., L. Braghieri, S. Eichmeyer, and M. Gentzkow (2020, March). The Welfare Effects of Social Media. *American Economic Review 110*(3), 629–76.

Allcott, H. and M. Gentzkow (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives 31*(2), 211–36.

Ambrosino, A., M. Cedrini, J. B. Davis, S. Fiori, M. Guerzoni, and M. Nuccio (2018). What Topic Modeling Could Reveal About the Evolution of Economics. *Journal of Economic Methodology 25*(4), 329–348.

Andres, R. and O. Slivko (2021). Combating Online Hate Speech: The Impact of Legislation on Twitter. *ZEW-Centre for European Economic Research Discussion Paper* (21-103).

Angelov, D. (2020). Top2vec: Distributed Representations of Topics. *arXiv preprint arXiv:2008.09470*.

Anti-Defamation League (2022). Online Hate and Harassment. The American Experience 2022. *Center for Technology and Society*. Accessed: 2022-09-11.

Aridor, G., R. Jiménez Durán, R. Levy, and L. Song (2024). The Economics of Social Media. *Available at SSRN*.

Ash, E. and S. Hansen (2023). Text Algorithms in Economics. *Annual Review of Economics 15*, 659–688.

Barrera, O., S. Guriev, E. Henry, and E. Zhuravskaya (2020). Facts, Alternative Facts, and Fact Checking In Times of Post-truth Politics. *Journal of Public Economics 182*(C).

Beknazar-Yuzbashev, G., R. Jiménez Durán, J. McCrosky, and M. Stalinski (2022). Toxic Content and User Engagement on Social Media: Evidence From a Field Experiment. *Available at SSRN*.

Beknazar-Yuzbashev, G., R. Jiménez Durán, and M. Stalinski (2024). A Model of Harmful yet Engaging Content on Social Media. *Available at SSRN*.

Bhuller, M., T. Havnes, E. Leuven, and M. Mogstad (2013). Broadband Internet: An Information Superhighway to Sex Crime? *Review of Economic Studies 80*(4), 1237–1266.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research 3*(Jan), 993–1022.

Bond, R. M., C. J. Fariss, J. J. Jones, A. D. Kramer, C. Marlow, J. E. Settle, and J. H. Fowler (2012). A 61-Million-Person Experiment in Social Influence and Political Mobilization. *Nature 489*(7415), 295.

Boxell, L., M. Gentzkow, and J. M. Shapiro (2017). Greater Internet Use Is Not Associated with Faster Growth in Political Polarization Among US Demographic Groups. *Proceedings of the National Academy of Sciences 114*(40), 10612–10617.

Braghieri, L., R. Levy, and A. Makarin (2022, November). Social Media and Mental Health. *American Economic Review 112*(11), 3660–3693.

Bundeskartellamt (2019). Bundeskartellamt Prohibits Facebook From Combining User Data From Different Sources. `https://www.bundeskartellamt.de/SharedDocs/Meldung/EN/Pressemitteilungen/2019/07_02_2019_Facebook.html`. Accessed: 2022-07-14.

Bursztyn, L., G. Egorov, R. Enikolopov, and M. Petrova (2019). Social Media and Xenophobia: Evidence from Russia. Working Paper 26567, National Bureau of Economic Research.

Bursztyn, L., G. Egorov, and S. Fiorin (2020). From Extreme to Mainstream: The Erosion of Social Norms. *American Economic Review 110*(11), 3522–48.

Campello, R. J., D. Moulavi, and J. Sander (2013). Density-Based Clustering Based on Hierarchical Density Estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pp. 160–172. Springer.

Cao, A., J. M. Lindo, and J. Zhong (2023). Can Social Media Rhetoric Incite Hate Incidents? Evidence From Trump's "Chinese Virus" Tweets. *Journal of Urban Economics 137*, 103590.

Card, D. and G. B. Dahl (2011). Family Violence and Football: The Effect of Unexpected Emotional Cues on Violent Behavior. *The Quarterly Journal of Economics 126*(1), 103–143.

Chen, Y. and D. Y. Yang (2019). The Impact of Media Censorship: 1984 or Brave New World? *American Economic Review 109*(6), 2294–2332.

Dahl, G. and S. DellaVigna (2009). Does Movie Violence Increase Violent Crime? *The Quarterly Journal of Economics*, 677–734.

De Chaisemartin, C. and X. d'Haultfoeuille (2023). Two-Way Fixed Effects and Differences-In-Differences With Heterogeneous Treatment Effects: A Survey. *The Econometrics Journal 26*(3), C1–C30.

DellaVigna, S., R. Enikolopov, V. Mironova, M. Petrova, and E. Zhuravskaya (2014, July). Cross-Border Media and Nationalism: Evidence from Serbian Radio in Croatia. *American Economic Journal: Applied Economics 6*(3), 103–32.

Deutscher Bundestag (2017). Drucksache 18/12356. `https://dserver.bundestag.de/btd/18/123/1812356.pdf`. Accessed: 2022-08-04.

Dixon, L., J. Li, J. Sorensen, N. Thain, and L. Vasserman (2018). Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 67–73.

Djourelova, M. (2023). Persuasion through Slanted Language: Evidence from the Media Coverage of Immigration. *American Economic Review 113*(3), 800–835.

Draca, M. and C. Schwarz (2024). How Polarized Are Citizens? Measuring Ideology From the Ground-up. *Forthcoming, Economic Journal*.

Echikson, W. and O. Knodt (2022). Germany's NetzDG: A Key Test for Combatting Online Hate. *Available at SSRN: https://ssrn.com/abstract=3300636*.

Economist (2018). In Germany, Online Hate Speech Has Real-World Consequences.

Enikolopov, R., A. Makarin, and M. Petrova (2020). Social Media and Protest Participation: Evidence from Russia. *Econometrica 88*(4), 1479–1514.

Enikolopov, R., M. Petrova, and K. Sonin (2018). Social Media and Corruption. *American Economic Journal: Applied Economics 10*(1), 150–174.

European Parliament (2016). European Parliament Resolution of 14 April 2016 on the 2015 Report on Turkey.

Fergusson, L. and C. Molina (2021, April). Facebook Causes Protests. Documentos CEDE 018002, Universidad de los Andes - CEDE.

Ferman, B., C. Pinto, and V. Possebom (2020). Cherry Picking With Synthetic Controls. *Journal of Policy Analysis and Management 39*(2), 510–532.

Fujiwara, T., K. MẌller, and C. Schwarz (2024). The Effect of Social Media on Elections: Evidence from The United States. *Journal of the European Economic Association 22*(3), 1495–1539.

Gentzkow, M., B. Kelly, and M. Taddy (2019). Text as Data. *Journal of Economic Literature 57*(3), 535–574.

Gillespie, T. (2018). *Custodians of the Internet*. Yale University Press.

Goebel, J., M. M. Grabka, S. Liebig, M. Kroh, D. Richter, C. Schröder, and J. Schupp (2019). The German Socio-Economic Panel (SOEP). *Jahrbcher fr Nationalökonomie und Statistik 239*(2), 345–360.

Gorwa, R. (2019). The Platform Governance Triangle: Conceptualising the Informal Regulation of Online Content. *Internet Policy Review 8*(2), 1–22.

Griffiths, T. L. and M. Steyvers (2004). Finding Scientific Topics. *Proceedings of the National Academy of Sciences 101*(suppl_1), 5228–5235.

Guriev, S., N. Melnikov, and E. Zhuravskaya (2021). 3G Internet and Confidence in Government. *The Quarterly Journal of Economics 136*(4), 2533–2613.

Habibi, M., D. Hovy, and C. Schwarz (2024). The Content Moderator's Dilemma: Removal of Toxic Content and Distortions to Online Discourse. *arXiv preprint arXiv:2412.16114*.

Hansen, S., M. McMahon, and A. Prat (2018). Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach. *The Quarterly Journal of Economics 133*(2), 801–870.

Heise (2017). "Facebook-Gesetz" in Kraft getreten.

Heldt, A. P. (2019). Reading Between the Lines and the Numbers: An Analysis of the First Netzdg Reports. *Internet Policy Review 8*(2).

Henry, E., S. Guriev, T. Marquis, and E. Zhuravskaya (2023, December). Curtailing False News, Amplifying Truth. CEPR Discussion Papers 18650, C.E.P.R. Discussion Papers.

Henry, E., E. Zhuravskaya, and S. Guriev (2022). Checking and Sharing Alt-facts. *American Economic Journal: Economic Policy 14*(3), 55–86.

Howard, P. N., A. Duffy, D. Freelon, M. Hussain, W. Mari, and M. Maziad (2011). Opening Closed Regimes: What Was the Role of Social Media During the Arab Spring? *Working Paper*.

Jiménez Durán, R. (2022). The Economics of Content Moderation: Theory and Experimental Evidence From Hate Speech on Twitter. *Available at SSRN*.

Jones, J. J., R. M. Bond, E. Bakshy, D. Eckles, and J. H. Fowler (2017, 04). Social Influence and Political Mobilization: Further Evidence From a Randomized Experiment in the 2012 U.S. Presidential Election. *PLOS ONE 12*(4), 1–9.

Justitia (2020). The Digital Berlin Wall: How Germany (Accidentally) Created a Prototype for Global Online Censorship - Act two.

Kaplan, J. (2025). More Speech and Fewer Mistakes. `https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/`.

Kaye, D. (2018). Mandate of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression. *OL ITA 1*(2018), 20.

Kaye, D. A. (2019). *Speech Police: The Global Struggle to Govern the Internet*. Columbia Global Reports.

Kohl, U. (2022). Platform Regulation of Hate Speech–A Transatlantic Speech Compromise? *Journal of Media Law*, 1–25.

Kominers, S. D. and J. M. Shapiro (2024). Content Moderation with Opaque Policies. Working Paper 32156, National Bureau of Economic Research.

Levy, R. (2021, March). Social Media, News Consumption, and Polarization: Evidence from a Field Experiment. *American Economic Review 111*(3), 831–70.

Liu, Y., P. Yildirim, and Z. J. Zhang (2021). Social Media, Content Moderation, and Technology. *arXiv preprint arXiv:2101.04618*.

Madio, L. and M. Quinn (2021). Content Moderation and Advertising in Social Media Platforms. *Available at SSRN 3551103*.

Manacorda, M. and A. Tesei (2020). Liberation Technology: Mobile Phones and Political Mobilization in Africa. *Econometrica 88*(2), 533–567.

𝕏 Safety (2023). Freedom of Speech, Not Reach: An update on our enforcement philosophy. `https://blog.x.com/en_us/topics/product/2023/freedom-of-speech-not-reach-an-update-on-our-enforcement-philosophy`.

Mosquera, R., M. Odunowo, T. McNamara, X. Guo, and R. Petrie (2020). The Economic Effects of Facebook. *Experimental Economics 23*(2), 575–602.

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 (2022). The Effects of Online Content Moderation: Evidence from President Trump's Account Deletion. *Available at SSRN 4296306*.

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

Netzpolitik (2019). Exklusiver Einblick: So funktionieren Facebooks Moderationszentren.

New York Times (2017). Seeking Asylum in Germany, and Finding Hatred, By Ainara Tiefenthäler, Shane O'neill and Andrew Michael Ellis .

Olden, A. and J. Møen (2022). The Triple Difference Estimator. *The Econometrics Journal 25*(3), 531–553.

Pemstein, D., K. L. Marquardt, E. Tzelgov, Y.-t. Wang, J. Krusell, and F. Miri (2018). The V-Dem Measurement Model: Latent Variable Analysis for Cross-National and Cross-Temporal Expert-Coded Data. *V-Dem Working Paper 21*.

Qin, B., D. Strömberg, and Y. Wu (2017). Why Does China Allow Freer Social Media? Protests Versus Surveillance and Propaganda. *Journal of Economic Perspectives 31*(1),

117–140.

Rauh, C. and L. Renée (2023). How to Measure Parenting Styles? *Review of Economics of the Household 21*(3), 1063–1081.

Reimers, N. and I. Gurevych (2019, 11). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Schwarz, C. (2023). Estimating Text Regressions Using txtreg_train. *The Stata Journal 23*(3), 799–812.

Stern (2018). Wie Facebook und Twitter das neue Löschgesetz umsetzen.

Sunstein, C. R. (2017). *# Republic: Divided Democracy in the Age of Social Media*. Princeton University Press.

Twitter (2015). Fighting Abuse to Protect Freedom of Expression. `https://blog.twitter.com/en_us/a/2015/fighting-abuse-to-protect-freedom-of-expression`. Accessed: 2022-09-11.

Twitter (2018a). Twitter Netzwerkdurchsetzungsgesetzbericht: Januar - Juni 2018.

Twitter (2018b). Twitter Netzwerkdurchsetzungsgesetzbericht: Juli - Dezember 2018.

Waldron, J. (2012). *The Harm in Hate Speech*. Cambridge, MA: Harvard University Press.

Wulczyn, E., N. Thain, and L. Dixon (2017). Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th international conference on world wide web*, pp. 1391–1399.

Yanagizawa-Drott, D. (2014). Propaganda and Conflict: Evidence from the Rwandan Genocide. *The Quarterly Journal of Economics 129*(4), 1947–1994.

Zeit (2018). Facebook will mit 10.000 neuen Mitarbeitern gegen Hetze vorgehen.

Zeit (2019). Deutsche Behörde verhängt Millionenstrafe gegen Facebook.

Zhuravskaya, E., M. Petrova, and R. Enikolopov (2020). Political Effects of the Internet and Social Media. *Annual Review of Economics 12*.

The Online and Offline Effects of Content Moderation: Evidence from Germany's NetzDG

Online Appendix

This Online Appendix consists of four parts.

1. Appendix A provides additional details on the data sources
2. Appendix B provides a theoretical framework for the empirical analysis.
3. Appendix C presents additional results on the online effects of the NetzDG.
4. Appendix D presents additional results on the offline effects of the NetzDG.

A Additional Details on the Data

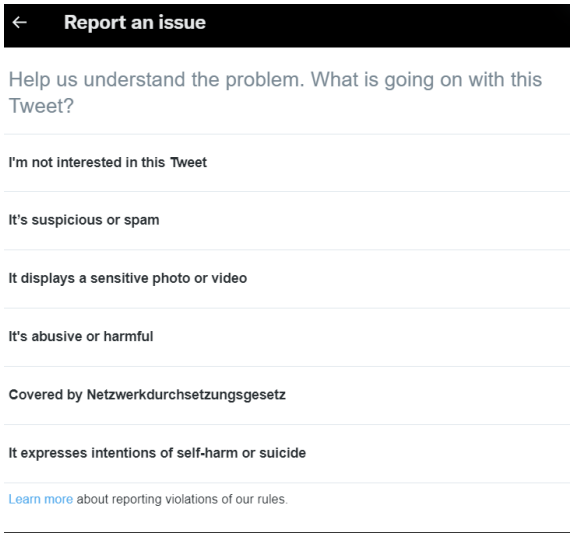
Table A.1: Translated Example of Toxic Refugee Tweets

Date	Post	Toxcity
2016-03-08	@{user} Oh {expletive} you (you refugee, go back to your country)	0.99
2020-09-12	@{user} @{user} I burn your pets like Hitler burned the fat corps of your ugly, {stream of expletives} grandparents in the concentration camps after they showered. You {expletive} refugee.	0.99
2017-11-19	@{user} It is normal. Every piece of trash that hurts, rapes and murders is given more attention than the victim. even more so when a fucking refugee. The police are instructed not to intervene so harshly	0.89
2018-03-07	@{user} What the fuck. I didn't vouch for any refugee. And now I have to pay for the stupidity of the do-gooders with my taxes. I find it an impudence.	0.86
2016-02-04	@{user} You're not a refugee otherwise, it would be free ;) the stupid German pays.	0.84

Notes: This table reports five examples of toxic refugee tweets. The tweets were translated by the authors. Usernames, expletives, and links were masked.

Figure A.1: How Twitter Users Can Report Content Covered by the NetzDG

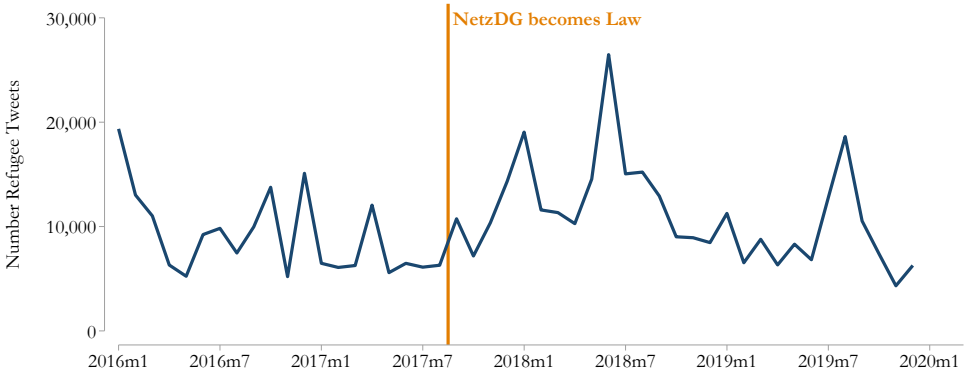
(a) Main Reporting Field



(b) Reason for Reporting

Notes: These screenshots show how Twitter users located in Germany can report content violating the NetzDG. Panel (a) shows the main reporting field a user sees when clicking on "report an issue" for a given tweet. Note that "Covered by the Netzwerkdurchsuchungsgesetz" is its own category. Panel (b) shows that the next prompt requires the user to specify a category, where "Hass schürende/verfassungswidrige Inhalte", "Gewalt/Bedrohung/Aufforderung zu Straftaten", "Beleidigung/Üble Nachrede", and "Terrorismus" refer directly to online hate speech or incitement of violence.

Figure A.2: Time Series Refugee Tweets



Notes: The time-series plot shows the monthly number of tweets mentioning the word "Flüchtling" (refugee) between 2016 and 2019.

Table A.2: Summary Statistics Toxicity Refugee Tweets

Variable	Mean	SD	p50	Min	Max	N
Toxicity Measures						
Toxicity	0.33	0.19	0.00	0.31	0.99	86,208
Sev. Toxicity	0.22	0.18	0.00	0.17	0.99	86,208
Identity Attack	0.46	0.22	0.00	0.46	1.00	86,208
Insult	0.29	0.18	0.00	0.26	1.00	86,208
Profanity	0.16	0.16	0.00	0.10	1.00	86,208
Threat	0.29	0.19	0.00	0.20	0.99	86,208
User Measures						
Pre-Period Tox $\geq$ 75pct	0.20	0.40	0.00	0.00	1.00	86,208
AfD User Pre-Period Tox $\geq$ 75pct	0.02	0.15	0.00	0.00	1.00	86,208

Notes: This table displays the mean, standard deviation, median, minimum, maximum, and number of observations for the variables used in the tweet-level analysis.

Table A.3: Examples of Anti-Refugee Incidents

Date	Place	Description	Type
15.06.2018	Ismaning	Two 28-year-old Germans met a group of people from Eritrea on the train. At first, the two made racist comments about them. After getting off at Ismaning train station, one of the two 28-year-olds pulled a 21-year-old from the group to the ground and kicked him. The injured person lost consciousness and had to be treated in a hospital.	Assault
27.09.2018	Werdau	A 25-year-old is said to have thrown an incendiary device onto the grounds of an asylum accommodation. Half an hour before the crime, the man threatened residents and security staff of the shelter that he would set the facility and the people living there on fire. He then left the crime scene and returned with the incendiary device to throw it over the entrance gate.	Arson
20.03.2016	Steinhagen	Garbage containers under the carport of an asylum accommodation catch fire for an unknown reason. Two garbage cans burn out completely in the fire.	Suspected Case
02.07.2016	Zirndorf	25 neo-Nazis from the alliance "Franken wehrt sicht" demonstrated in the afternoon under the motto "Zirndorf says no to the home - citizen dialogue now!" in front of an asylum accommodation.	Demonstration
01.09.2018	Leipzig	Two masked men riot with a baseball bat and a pool cue in front of the house where a 31-year-old asylum seeker lives with his wife and five children.	Other cases

Notes: This table reports one example for each class of anti-refugee incidents in the data. The descriptions were translated by the authors.

Table A.4: Summary Statistics

Variable	Mean	SD	p50	Min	Max	N
Anti-Refugee Incidents						
Anti-refugee incidents	0.14	1.07	0.00	0.00	115.00	71,456
Anti-refugee incidents (arson)	0.00	0.06	0.00	0.00	9.00	71,456
Anti-refugee incidents (demonstration)	0.00	0.04	0.00	0.00	4.00	71,456
Anti-refugee incidents (assault)	0.02	0.23	0.00	0.00	15.00	71,456
Anti-refugee incidents (other)	0.11	0.86	0.00	0.00	88.00	71,456
Anti-refugee incidents (suspected cases)	0.00	0.11	0.00	0.00	13.00	71,456
Main Variables						
AfD users per capita (in %)	0.03	0.02	0.00	0.03	0.11	71,456
Log(Population)	9.15	0.93	5.81	9.10	15.07	71,456
Vote share AfD	14.86	7.01	3.13	12.85	44.86	71,008
Facebook users per capita	0.08	0.12	0.00	0.05	0.91	71,456
Share broadband internet (in %)	83.00	10.66	43.50	84.60	100.00	71,456
Additional Control Variables						
GDP per worker	63094.77	9846.31	46835.00	62207.00	136763.00	71,152
Population density	281.92	381.64	6.55	144.77	4653.18	71,456
Immigrants per capita	13.96	7.63	1.82	13.78	49.72	69,632
Refugees per capita	0.01	0.01	0.00	0.01	0.10	71,456
Registered domains per capita	0.14	0.06	0.06	0.13	1.39	71,456
Mobile broadband speed	11.90	2.33	6.24	11.60	24.41	71,456
Newspaper sales per capita	0.09	0.08	0.00	0.09	1.64	70,800
Vote share CDU	36.45	7.10	19.88	35.74	64.48	71,008
Vote share SPD	18.55	7.04	4.68	17.23	46.70	71,008
Vote share Linke	7.84	4.37	1.57	6.16	26.10	71,008
Vote share Green	7.03	3.50	0.87	6.66	25.47	71,008
Vote share FDP	9.70	2.87	3.38	9.29	27.52	71,008
Vote share NPD	0.49	0.41	0.00	0.31	2.01	71,456
Voter turnout	76.44	3.14	65.93	76.46	83.88	71,456
Average age	44.97	2.28	26.80	44.70	56.20	69,168
Share population 0-25	24.73	3.18	13.78	25.19	37.14	69,168
Share population 25-50	33.35	2.04	21.67	33.32	45.37	69,168
Share population 50-75	32.58	3.14	21.97	32.14	50.08	69,168
Share population 75+	9.34	1.81	3.58	9.22	17.65	69,168

Notes: This table displays the mean, standard deviation, median, minimum, maximum, and number of observations of the variables used in the municipality-quarter panel.

Table A.5: Summary Statistics by Quartile of AfD Facebook Users Per Capita

	1st Quartile		2nd Quartile		3rd Quartile		4th Quartile	
Variable	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Anti-refugee incidents	0.041	0.256	0.077	0.374	0.114	0.466	0.332	2.023
Anti-refugee incidents (arson)	0.000	0.022	0.002	0.050	0.002	0.050	0.004	0.094
Anti-refugee incidents (demonstration)	0.000	0.011	0.000	0.021	0.000	0.021	0.003	0.073
Anti-refugee incidents (assault)	0.004	0.084	0.009	0.111	0.017	0.157	0.065	0.415
Anti-refugee incidents (other)	0.034	0.222	0.063	0.325	0.090	0.386	0.250	1.617
Anti-refugee incidents (suspected cases)	0.001	0.044	0.003	0.069	0.004	0.111	0.011	0.171
AfD users per capita (in %)	0.002	0.004	0.019	0.004	0.034	0.005	0.063	0.018
Log(Population)	8.605	0.728	9.287	0.630	9.370	0.875	9.357	1.170
Vote share AfD	14.665	6.828	13.480	6.153	14.663	6.731	16.645	7.848
Facebook users per capita	0.064	0.121	0.084	0.131	0.086	0.115	0.086	0.098
Share broadband internet (in %)	82.737	9.859	83.633	10.184	83.196	11.256	82.433	11.215
GDP per worker	63297.784	9717.812	63976.647	10014.253	63726.393	9901.268	61373.485	9532.920
Population density	202.268	293.691	261.068	306.674	314.564	385.356	349.824	491.318
Immigrants per capita	12.913	6.617	15.095	7.253	15.016	7.726	12.837	8.495
Refugees per capita	0.010	0.005	0.011	0.005	0.011	0.007	0.011	0.007
Registered domains per capita	0.142	0.055	0.143	0.048	0.142	0.049	0.138	0.069
Mobile broadband speed	11.737	2.321	11.855	2.389	11.937	2.296	12.064	2.300
Newspaper sales per capita	0.117	0.085	0.086	0.071	0.083	0.071	0.084	0.073
Vote share CDU	38.718	7.284	37.010	6.760	35.746	6.635	34.311	6.968
Vote share SPD	17.033	6.751	19.426	7.012	19.450	6.848	18.288	7.251
Vote share Linke	6.809	3.916	7.303	3.810	7.865	4.162	9.381	5.060
Vote share Green	7.146	3.569	7.512	3.400	7.023	3.320	6.447	3.636
Vote share FDP	9.344	2.826	10.172	2.884	10.020	3.007	9.270	2.659
Vote share NPD	0.468	0.387	0.425	0.356	0.475	0.397	0.597	0.471
Voter turnout	76.904	3.006	76.836	2.980	76.368	3.057	75.669	3.333
Average age	44.687	2.301	44.621	2.069	44.980	2.119	45.608	2.465
Share population 0-25	25.294	3.170	25.326	2.970	24.672	2.957	23.624	3.307
Share population 25-50	33.519	2.017	33.496	1.885	33.343	1.923	33.050	2.267
Share population 50-75	32.236	3.149	32.116	2.915	32.588	2.919	33.378	3.393
Share population 75+	8.951	1.791	9.062	1.639	9.397	1.716	9.948	1.921

Notes: This table displays the mean, standard deviation, of the variables used in the municipality-year-quarter panel, split by quartiles of AfD Facebook users per capita (the "exposure" variable in the difference-in-differences analysis).

Table A.6: Summary Statistics GSOEP

Variable	Mean	SD	p50	Min	Max	N
Pro-refugee Attitudes						
Index refugee attitudes	0.49	0.38	0.00	0.60	1.00	36,912
Refugees are good for economy	0.59	0.49	0.00	1.00	1.00	36,682
Refugees are good for culture	0.56	0.50	0.00	1.00	1.00	36,677
Refugees are good for place of living	0.52	0.50	0.00	1.00	1.00	36,665
Refugee are a chance (Short-term)	0.25	0.44	0.00	0.00	1.00	36,667
Refugee are a chance (Long-term)	0.52	0.50	0.00	1.00	1.00	36,598
Pro-refugee Actions						
Index refugee actions	0.15	0.23	0.00	0.00	1.00	36,979
Action: Donated (Last Year)	0.27	0.44	0.00	0.00	1.00	36,861
Action: Donated (Future)	0.30	0.46	0.00	0.00	1.00	36,332
Action: Volunteered (Last Year)	0.07	0.25	0.00	0.00	1.00	36,761
Action: Volunteered (Future)	0.11	0.31	0.00	0.00	1.00	36,225
Action: Demonstrated (Last Year)	0.05	0.21	0.00	0.00	1.00	36,733
Action: Demonstrated (Future)	0.08	0.27	0.00	0.00	1.00	36,202
Respondent Characteristics						
Social media user	0.63	0.48	0.00	1.00	1.00	41,644
AfD voter	0.06	0.24	0.00	0.00	1.00	41,644
Female	0.53	0.50	0.00	1.00	1.00	41,644
Age	49.23	17.11	18.00	48.00	103.00	41,643

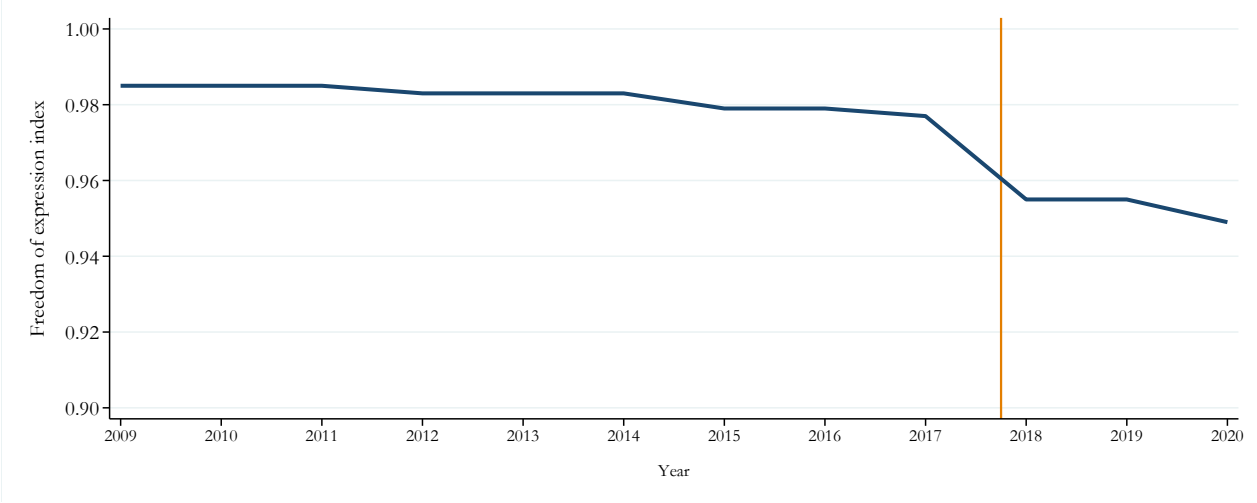
Notes: This table displays the mean, standard deviation, median, minimum, maximum, and number of observations for the variables from GSOEP used in the attitudes analysis.

Table A.7: OSCE Members and Data Filters

OSCE State	No data 2009-2020	Microstate	Data changes 2017-2020	7+ missings 2009-2020	End gaps
Albania				×	×
Andorra		×			
Armenia				×	×
Austria					
Azerbaijan				×	×
Belarus				×	×
Belgium					
Bosnia and Herzegovina					
Bulgaria					
Canada			×		
Croatia					
Cyprus					
Czech Republic					
Denmark					
Estonia				×	
Finland					
France					×
Georgia			×		
Germany					
Greece			×		
Holy See		×			×
Hungary			×		
Iceland					×
Ireland			×		
Italy					
Kazakhstan					×
Kyrgyzstan				×	×
Latvia					×
Liechtenstein		×			
Lithuania					
Luxembourg	×			×	×
Malta	×	×		×	×
Moldova					
Monaco	×			×	×
Mongolia					×
Montenegro				×	×
Netherlands			×		
North Macedonia				×	×
Norway			×		
Poland					
Portugal					
Romania				×	×
Russian Federation				×	×
San Marino	×	×		×	×
Serbia			×		
Slovakia					
Slovenia			×	×	
Spain					
Sweden			×		
Switzerland					
Tajikistan	×			×	×
Turkey					
Turkmenistan	×			×	×
UK					
US					
Ukraine					
Uzbekistan				×	×

Notes: This table presents the list of the 57 OSCE member States and the selection criteria used to filter them. Germany and the donors in the baseline specification are bolded. "No data 2009-2020" indicates that there was no data for that period. "Microstate" indicates microstates. "End gaps" indicates missing data at the beginning or end of the series, even after interpolation (i.e., countries that would require extrapolation to be balanced). "7+ missings 2009-2020" indicates that the raw data has more than 7 years of missing values. "Data changes 2017-2020" indicates changes in the measurement of hate crimes in that period.

Figure A.3: Freedom of Expression Index, 2009-2020



Notes: This graph shows the Freedom of Expression Index for Germany obtained from the V-Dem dataset (Pemstein et al., 2018). This index aggregates the ratings provided by multiple country experts who respond to questions regarding government censorship efforts, the harassment of journalists, media self-censorship, media bias, freedom of discussion for ordinary citizens, and freedom of academic and cultural expression. For reference, the mean value of the index pre-NetzDG across countries was 0.68, and the standard deviation was 0.28.

B Theoretical Framework

This model builds on the microfoundation laid out in Jiménez Durán (2022). The model assumes that there is a single platform on which two types of users—"Acceptable" (A) and "Hater" (H)—interact with each other. The platform chooses a moderation rate $c \in [0, 1]$ that determines the proportion of hateful content that survives on the platform. Moreover, by carefully choosing its advertising frequencies, the platform can effectively choose the engagement of each type of user; that is, the amount of time they spend consuming content. Let T^A denote the aggregate engagement of acceptable users and T^H denote the aggregate engagement of hateful users post-moderation.

The platform faces inverse demands $p^\theta(T^A, T^H, c)$, $\theta \in \{A, H\}$. These objects equal the amount of dollars that advertisers are willing to pay per minute of ad times the amount of time that users are willing to spend watching ads per minute of content consumed.[29] The platform also has costs $\phi(T^A, T^H, c)$ and is required by a regulator to pay an expected penalty $\tau > 0$ for each unit of hateful content that it fails to moderate. Hence, its problem becomes:

$$\max_{T^A, T^H, c} \; p^A(T^A, T^H, c)T^A + p^H(T^A, T^H, c)T^H - \phi(T^A, T^H, c) - \tau T^H. \tag{B.1}$$

We interpret the implementation of the NetzDG as a marginal increase in the expected regulatory penalty; $d\tau > 0$. In other words, the policy resulted in an increase in the marginal cost of unmoderated hate speech. In this case, it is easy to show that, if the second-order conditions of problem (B.1) hold, the amount of surviving hateful content on the platform decreases in response to an increase in fines; $dT^H/d\tau < 0$.[30]

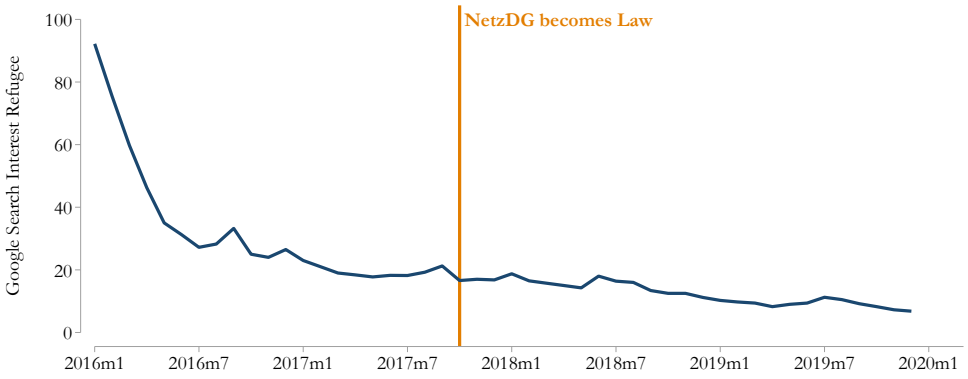
[29] In the notation of Jiménez Durán (2022), $p^\theta(T^A, T^H, c) = a^\theta(T^A, T^H, c)P^\theta(T^A, T^H, c)$, where a^θ denotes the advertisers' willingness to pay and P^θ denotes the advertising load for type θ. In this paper, we allow the platform to be a price-setter in the ads market.

[30] To see why, rewrite problem (B.1) as $\max_{T^H} \; \widetilde{\pi}(T^H) - \tau T^H$, where $\widetilde{\pi}(T^H)$ denotes the maximized profits (pre-penalties) for a given T^H. Applying the implicit function theorem to the first-order condition of this problem yields $dT^H/d\tau = 1/\widetilde{\pi}''$. The second-order condition of the problem requires that $\widetilde{\pi}'' < 0$.

C Additional Results on Online Effects

C.1. Additional Results for the Toxicity of Refugee Tweets

Figure C.1: Time Series Google Trends Refugee



Notes: The time-series plot shows the monthly Google search interest for the "Flüchtling" topic in Germany between 2016 and 2019.

Table C.1: Robustness: Threshold of Pre-Period Toxicity

	Dep. var.: Toxicity Refugee Tweets			
	(1)	(2)	(3)	(4)
Panel (a): Highly Toxic Users				
Toxic User (≥50pct) × Post	-0.123***			
	(0.003)			
Toxic User (≥75pct) × Post		-0.166***		
		(0.004)		
Toxic User (≥90pct) × Post			-0.261***	
			(0.006)	
Toxic User (≥95pct) × Post				-0.337***
				(0.008)
User FE	Yes	Yes	Yes	Yes
Quarter FE	Yes	Yes	Yes	Yes
Observations	86,208	86,208	86,208	86,208
Pre-Period Mean of DV	0.31	0.31	0.31	0.31
R^2	0.44	0.44	0.44	0.44
Panel (b): Toxic AfD Followers				
Toxic AfD User (≥50pct) × Post	-0.048***			
	(0.005)			
Toxic AfD User (≥75pct) × Post		-0.104***		
		(0.009)		
Toxic AfD User (≥90pct) × Post			-0.250***	
			(0.021)	
Toxic AfD User (≥95pct) × Post				-0.356***
				(0.025)
User FE	Yes	Yes	Yes	Yes
Quarter FE	Yes	Yes	Yes	Yes
Observations	86,208	86,208	86,208	86,208
Pre-Period Mean of DV	0.31	0.31	0.31	0.31
R^2	0.43	0.43	0.43	0.43

Notes: This table presents the results of estimating Equation (1) where the dependent variable is the toxicity of tweets containing the word "Flüchtling" (refugee) (bounded between 0 and 1). In panel (a), $Toxic\ User_i$ is an indicator variable equal to 1 if a users' refugee tweets before the NetzDG were on average above the 50th, 75th, 90th, or 95th percentile. In panel (b) *Toxic AfD follower* is an indicator variable that is equal to 1 if a Twitter user additionally follows the AfD's account. All regressions control for user and day fixed effects. Robust standard errors in parentheses are clustered by user. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table C.2: Robustness: Toxicity Measures – Refugee Tweets

	Dep. var.: Toxicity measured by:					
	Toxicity	Severe Toxicity	Identity Attack	Insult	Profanity	Threat
	(1)	(2)	(3)	(4)	(5)	(6)
Panel (a): Toxic Users						
Toxic User × Post	-0.166***	-0.152***	-0.174***	-0.150***	-0.117***	-0.098***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Quarter FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	86,208	86,208	86,208	86,208	86,208	86,208
Pre-Period Mean of DV	0.31	0.21	0.44	0.28	0.15	0.28
R^2	0.44	0.40	0.44	0.45	0.38	0.35
Panel (b): Toxic AfD Followers						
Toxic AfD User × Post	-0.104***	-0.092***	-0.098***	-0.097***	-0.074***	-0.047***
	(0.009)	(0.011)	(0.011)	(0.010)	(0.010)	(0.011)
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Quarter FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	86,208	86,208	86,208	86,208	86,208	86,208
Pre-Period Mean of DV	0.31	0.21	0.44	0.28	0.15	0.28
R^2	0.43	0.38	0.42	0.44	0.36	0.34

Notes: This table presents the results of estimating Equation (1), where the dependent variable is the measure of toxicity listed in the top row, bounded between 0 and 1, calculated based on tweets containing the word refugee ("Flüchtling"). In panel (a), we use and indicator variable equal to 1 if a users' refugee tweets before the NetzDG were on average above the 75th percentile. In panel (b) *Toxic AfD follower* is an indicator variable that is equal to 1 if a Twitter user additionally follows the AfD's account. All regressions control for AfD follower and day fixed effects. Robust standard errors in parentheses are clustered by users. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table C.3: Robustness: Threshold of Maximum Pre-Period Toxicity

	Dep. var.: Toxicity of Refugee Tweets			
	(1)	(2)	(3)	(4)
Toxic User ($\max(tox) \geq 0.2$) × Post	-0.191***			
	(0.005)			
Toxic User ($\max(tox) \geq 0.3$) × Post		-0.172***		
		(0.004)		
Toxic User ($\max(tox) \geq 0.4$) × Post			-0.136***	
			(0.003)	
Toxic User ($\max(tox) \geq 0.5$) × Post				-0.108***
				(0.003)
User FE	Yes	Yes	Yes	Yes
Quarter FE	Yes	Yes	Yes	Yes
Observations	86,208	86,208	86,208	86,208
Pre-Period Mean of DV	0.31	0.31	0.31	0.31
R^2	0.44	0.44	0.44	0.44

Notes: This table presents the results of estimating Equation (1) where the dependent variable is the toxicity of tweets containing the word "Flüchtling" (refugee) (bounded between 0 and 1). $Toxic\ User_i$ is an indicator variable equal to 1 if a users' refugee tweets before the NetzDG posted a tweet with a toxicity above 0.2, 0.3, 0.4, or 0.5. All regressions control for user and day fixed effects. Robust standard errors in parentheses are clustered by user. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

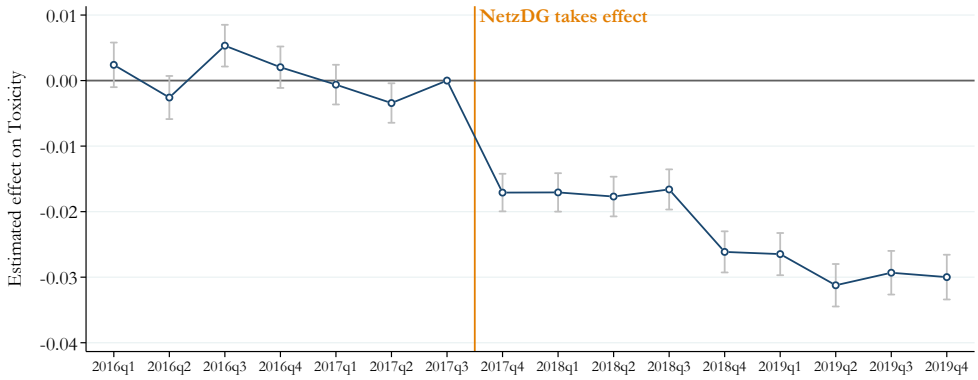
Table C.4: NetzDG and Refugee-related Twitter Activity

Panel (a):	*Dep. var.: Asinh(Nr. Refugee Tweets)*			
	(1)	(2)	(3)	(4)
Toxic User × Post	0.161***	-0.015		
	(0.006)	(0.009)		
Toxic AfD User × Post			0.022	0.069
			(0.039)	(0.052)
User FE	Yes	Yes	Yes	Yes
Quarter FE	Yes	Yes	Yes	Yes
Linear Time Trend		Yes		Yes
Observations	361,592	361,592	361,592	361,592
Pre-Period Mean of DV	0.64	0.64	0.64	0.64
R^2	0.45	0.58	0.45	0.58
Panel (b):	*Dep. var.: Nr. Refugee Tweets*			
	(1)	(2)	(3)	(4)
Toxic User × Post	0.591***	-0.097*		
	(0.058)	(0.053)		
Toxic AfD User × Post			-1.008	-0.164
			(0.744)	(0.585)
User FE	Yes	Yes	Yes	Yes
Quarter FE	Yes	Yes	Yes	Yes
Linear Time Trend		Yes		Yes
Observations	361,592	361,592	361,592	361,592
Pre-Period Mean of DV	1.73	1.73	1.73	1.73
R^2	0.41	0.75	0.41	0.75

Notes: This table presents the results of estimating from two-way fixed effect regression in a balanced panel of Twitter users, where the dependent variable in panel (a) is the inverse hyperbolic sine of the number of tweets containing the word "Flüchtling" (refugee) send by user i in quarter t. Panel (b) shows the same specification for the tweet counts. In columns (1) and (2), $Toxic\,User_i$ is an indicator variable equal to 1 if a user's refugee tweets before the NetzDG were, on average, above the 75th percentile. In columns (3) and (4), $AfDfollowers_i$ is an indicator variable that is equal to 1 if a Twitter user follows the AfD's account. All regressions control for user and quarter fixed effects. Columns (2) and (4) additionally control for user-specific linear time trends. Robust standard errors in parentheses are clustered by user. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

C.2. Additional Results for the Toxicity of All Tweets

Figure C.2: NetzDG and Overall Online Toxicity



Notes: The figure plots the coefficients from event study versions of Equation (1). The dependent variable is the average toxicity of all tweets sent by the users from our main analysis. The omitted category is the 3rd quarter of 2017, the quarter before the passing of the NetzDG (indicated with the vertical line). The whiskers indicate 95% confidence intervals based on standard errors clustered by user.

Table C.5: Regression Estimates: NetzDG and Overall Online Toxicity

	Dep. var.: Toxicity of All Tweets			
	(1)	(2)	(3)	(4)
Toxic User × Post	-0.023***	-0.014***		
	(0.001)	(0.001)		
Toxic AfD User × Post			-0.009***	-0.006*
			(0.002)	(0.003)
User FE	Yes	Yes	Yes	Yes
Quarter FE	Yes	Yes	Yes	Yes
Linear Time Trend		Yes		Yes
Observations	610,059	610,059	610,059	610,059
Pre-Period Mean of DV	0.17	0.17	0.17	0.17
R^2	0.55	0.63	0.55	0.63

Notes: This table presents the results of estimating from two-way fixed effect regression in a balanced panel of Twitter users, where the dependent variable is either the average toxicity of tweets (columns 1 and 2) or the inverse hyperbolic sine of the number of tweets (columns 3 and 4) send by user i in quarter t. $Toxic\ User_i$ is an indicator variable equal to 1 if a user's tweets before the NetzDG were, on average, above the 75th percentile. $Toxic\ AfD\ User_i$ is an indicator variable equal to 1 if a user's tweets before the NetzDG were, on average, above the 75th percentile and the user followed the AfD. All regressions control for user and quarter fixed effects. Columns (2) and (4) additionally control for user-specific linear time trends. Robust standard errors in parentheses are clustered by user. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table C.6: Robustness: Threshold of Pre-Period Toxicity

	Dep. var.: Toxicity of All Tweets			
	(1)	(2)	(3)	(4)
Panel (a): Toxic Users				
Pre-Period Tox $\geq$ 50pct $\times$ Post	-0.019***			
	(0.001)			
Pre-Period Tox $\geq$ 75pct $\times$ Post		-0.023***		
		(0.001)		
Pre-Period Tox $\geq$ 90pct $\times$ Post			-0.036***	
			(0.001)	
Pre-Period Tox $\geq$ 95pct $\times$ Post				-0.048***
				(0.002)
User FE	Yes	Yes	Yes	Yes
Quarter FE	Yes	Yes	Yes	Yes
Observations	610,059	610,059	610,059	610,059
Pre-Period Mean of DV	0.17	0.17	0.17	0.17
R^2	0.55	0.55	0.55	0.55
Panel (b): Toxic AfD Followers				
Toxic AfD User ($\geq$50pct) $\times$ Post	-0.003*			
	(0.002)			
Toxic AfD User ($\geq$75pct) $\times$ Post		-0.009***		
		(0.002)		
Toxic AfD User ($\geq$90pct) $\times$ Post			-0.018***	
			(0.003)	
Toxic AfD User ($\geq$95pct) $\times$ Post				-0.026***
				(0.005)
User FE	Yes	Yes	Yes	Yes
Quarter FE	Yes	Yes	Yes	Yes
Observations	610,059	610,059	610,059	610,059
Pre-Period Mean of DV	0.17	0.17	0.17	0.17
R^2	0.55	0.55	0.55	0.55

Notes: This table presents the results of estimating Equation (1) where the dependent variable is the toxicity of tweets containing the word "Flüchtling" (refugee) (bounded between 0 and 1). In panel (a), $Toxic\,User_i$ is an indicator variable equal to 1 if a users' tweets before the NetzDG were on average above the 50th, 75th, 90th, or 95th percentile. In panel (b), we additionally restrict to users who toxic users who follow the AfD. All regressions control for user and day fixed effects. Robust standard errors in parentheses are clustered by user. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table C.7: Robustness: Toxicity Measures – All Tweets

	Dep. var.: Toxicity measured by:					
	Toxicity	Severe Toxicity	Identity Attack	Insult	Profanity	Threat
	(1)	(2)	(3)	(4)	(5)	(6)
Panel (a): Toxic Users						
Toxic User × Post	-0.023***	-0.017***	-0.015***	-0.021***	-0.014***	-0.011***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Quarter FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	610,059	610,059	610,059	610,059	610,059	610,059
Pre-Period Mean of DV	0.17	0.10	0.14	0.15	0.11	0.16
R^2	0.55	0.51	0.58	0.58	0.46	0.59
Panel (b): Toxic AfD Followers						
Toxic AfD User × Post	-0.009***	-0.004**	-0.002	-0.006***	-0.004*	-0.001
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Quarter FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	610,059	610,059	610,059	610,059	610,059	610,059
Pre-Period Mean of DV	0.17	0.10	0.14	0.15	0.11	0.16
R^2	0.55	0.51	0.58	0.58	0.46	0.59

Notes: This table presents the results of estimating Equation (1), where the dependent variable is the measure of toxicity listed in the top row, bounded between 0 and 1, calculated based on tweets containing the word refugee ("Flüchtling"). In panel (a), we use and indicator variable equal to 1 if a users' tweets before the NetzDG were on average above the 75th percentile. In panel (b) *AfD follower* is an indicator variable that is equal to 1 if a users' tweets before the NetzDG were on average above the 75th percentile and if a user follows the AfD's account. All regressions control for AfD follower and day fixed effects. Robust standard errors in parentheses are clustered by users. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table C.8: Robustness: Threshold of Maximum Pre-Period Toxicity

	Dep. var.: Toxicity of All Tweets			
	(1)	(2)	(3)	(4)
Toxic User ($\max(tox) \geq 0.2$) × Post	-0.084***			
	(0.003)			
Toxic User ($\max(tox) \geq 0.3$) × Post		-0.066***		
		(0.002)		
Toxic User ($\max(tox) \geq 0.4$) × Post			-0.048***	
			(0.001)	
Toxic User ($\max(tox) \geq 0.5$) × Post				-0.034***
				(0.001)
User FE	Yes	Yes	Yes	Yes
Quarter FE	Yes	Yes	Yes	Yes
Observations	610,059	610,059	610,059	610,059
Pre-Period Mean of DV	0.17	0.17	0.17	0.17
R^2	0.55	0.55	0.55	0.55

Notes: This table presents the results of estimating Equation (1) where the dependent variable is the toxicity of tweets (bounded between 0 and 1). $Toxic\,User_i$ is an indicator variable equal to 1 if a users' tweets before the NetzDG posted a tweet with a toxicity above 0.2, 0.3, 0.4, or 0.5. All regressions control for user and day fixed effects. Robust standard errors in parentheses are clustered by user. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table C.9: NetzDG and Twitter Activity

Panel (a):	*Dep. var.: Asinh(Nr. Tweets)*			
	(1)	(2)	(3)	(4)
Toxic User × Post	0.094***	0.004		
	(0.015)	(0.015)		
Toxic AfD User × Post			0.122***	0.082*
			(0.043)	(0.045)
User FE	Yes	Yes	Yes	Yes
Quarter FE	Yes	Yes	Yes	Yes
Linear Time Trend		Yes		Yes
Observations	627,610	627,610	627,610	627,610
Pre-Period Mean of DV	4.28	4.28	4.28	4.28
R^2	0.54	0.74	0.54	0.74
Panel (b):	*Dep. var.: Nr. Tweets*			
	(1)	(2)	(3)	(4)
Toxic User × Post	3.494***	-1.371		
	(1.243)	(1.515)		
Toxic AfD User × Post			-1.532	4.790
			(3.669)	(4.644)
User FE	Yes	Yes	Yes	Yes
Quarter FE	Yes	Yes	Yes	Yes
Linear Time Trend		Yes		Yes
Observations	627,610	627,610	627,610	627,610
Pre-Period Mean of DV	4.28	4.28	4.28	4.28
R^2	0.43	0.66	0.43	0.66

Notes: This table presents the results of estimating from two-way fixed effect regression in a balanced panel of Twitter users, where the dependent variable in panel (a) is the inverse hyperbolic sine of the number of tweets send by user i in quarter t. Panel (b) shows the same specification for the tweet counts. In columns (1) and (2), $Toxic\,User_i$ is an indicator variable equal to 1 if a user's tweets before the NetzDG were, on average, above the 75th percentile. In columns (3) and (4), $Toxic\,AfD\,followers_i$ is an indicator variable that is equal to 1 if a Twitter user additionally follows the AfD's account. All regressions control for user and quarter fixed effects. Columns (2) and (4) additionally control for user-specific linear time trends. Robust standard errors in parentheses are clustered by user. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

C.3. Additional Results for Content Changes

Figure C.3: Changes in Word Frequencies – All Tweets

(a) Decreasing Frequency

(b) Increasing Frequency

Notes: This figure shows word clouds that visualize the relative word frequency changes for toxic compared to non-toxic users after the NetzDG among all tweets. Panel (a) shows words with decreasing relative frequency, while panel (b) shows words with increasing relative frequency. The size of the words is proportional to the frequency change.

Table C.10: Overview Topic Model: Falling Topics

Topic Nr.	Topic Label	Topic Words
67	Online Commentary	positive, positive, positive, positive, positive, negative, negative, negative, video, video evidence, negative, optimistic, optimism, strachevideo, commented, videos, youtube, music video, feedback, comments, comment, minus, comments, comment, hate comments, youtubers, youtuber, commentators, reaction, reactions, criticized, stability, review, commentator, criticize, reacts, comments, response, neutrality, react, satisfaction, neutral, ratings, rating, clip, clipmedianews, rated, viral, test, neutral
21	Turkey & Refugees	erdogan, erdogan, turkish, turkish, turk, turkish, turkish, turkey, turchen, turkey, turk, turkish, istanbul, turks, jihadists, gundogan, syrian, syrian, armenia, iranian, syrian, iran, kurdish, iranian, iranian, kurdish, iranian, kurds, taliban, iraqi, islamist, islamophobia, pakistan, refugees, islamists, syrians, asylum policy, islamist, asylum crisis, croatia, islamic, islamic, right-wing extremist, islamic, islamist, islamist, islamization, neo-nazis, constantine, refugees
15	Public Holidays & Event	new year reception, april, new year cleaning, new year, november, https, october, september, june, mars, previous year, april fool's joke, january, february, year of life, july, this year, august, anno, year, this year, decades, election night, decades, spring, the other day, start of the week, new beginning, oktoberfest, election year, today show, decades, zuckerberg, new elections, valentine's day, tomorrow, decade, berlin election, morning post, federal election, quarter finals, tomorrow, elections, inauguration, to begin, started, contemporary, debut, Reichstag, marcus
0	Informal Opinions	stabbing, princess, mess, for my sake, brewery, district office, amok, nogroko, me, imo, moi, nintendo, imam, sam, ergo, imho, lk, hmmm, yo, hahah, ahhh, pforzheim, ohhh, take, mi, hmm, ios, eu, bim, ahh, I, mine, kerstin, ohh, ohm, oh well, uff, ehm, hah, diego, uncomfortable, hahaha, hahahaha, my, hh, mia, haha, imm, unpleasant, ypg
41	Foreign Politics	obama, federal president, presidential election, obamas, republicans, president, wikileaks, candidate for chancellor, president, president, assange, anti-democrats, goes to the elections, chair, bolsonaro, chairman, democrat, americans, election night, democrats, liberal, new elections, women's suffrage, liberals, tweet, america, chairman, america, social democrat, erdogan, twitter, chairman, twitterer, zuckerberg, spd chairmanship, erdogan, protest voter, neoliberal, liberal, america, twitter account, tweet, neo-nazis, federal republic, american, tweet, neoliberalism, retweet, neoliberal, american
4	Sport	women's football, football, football fans, football game, handball, soccer, world cup, footballer, eurosport, national player, fcbayern, schweinsteiger, ltwbayern, hopesheim, women's football, ancelotti, champions league, sports director, stadium, league, cup final, olympics, sports show, hockey, teams, superbowl, ice hockey, European champions, semi-finals, Upper Bavaria, derby, sports, Lower Bavaria, round of 16, sporty, basketball, Olympics, sport, playing field, coach, athlete, esports, the team, DFB Cup, sports studio, club, team, playoffs, Bierhoff
62	Terrorism	terrorist, terrorist, terrorist attack, terrorism, terrorists, terrorist attacks, terrorist, terrorist group, terrorist militia, suspected terrorism, terror threat, terror, bomb attack, terrors, bombed, extremists, world war bomb, aerial bomb, bomb, bomb threat, extremism, right-wing terror, bombs, jihadists, assassin, assassination, right-wing extremist, mass murderer, bomber, atomic bomb, islamophobia, explosion, islamists, explode, explosions, mass murder, islamist, islamist, islamist, killer, death threats, deep, islamism, islamist, murder, murder, violent, attacks, buffet, murder case
34	Elections	elections, new elections, local elections, deselections, parliamentary election, presidential election, local election, select, protest voters, state elections, women's suffrage, election night, re-election, European elections, state election, voting, Berlin election, Bavarian election, word choice, votes, postal vote, eu election, candidate for chancellor, new election, vote, vote, run-off election, european election, non-voter, hessian election, selected, elected, voted out, referendum, selection, elected, choose, selected, voted, federal election, elected, democracy, candidates, democracies, democratic, democratic, undemocratic, run for office, free voter, democratic
19	Local Events	augsburg, wurzburg, harburg, freiburg, aschaffenburg, tecklenburg, stronghold, petersburg, neubrandenburg, stauffenberg, flensburg, homburg, wolfsburg, poggenburg, ravensburg, strasbourg, mecklenburg, wurttemberg, heidelberg, regensburg, ludwigsburg, luneburg, hamburg, brandenburg, magdeburg, oldenburg, hopenheim, salzburg, charlottenburg, lindenberg, duisburg, brandenburger, gretathunberg, wittenberg, vorarlberg, luxembourg, nurnberg, reinickendorf, hamburg, marburg, hambacherwald, train stations, ingolstadt, luxemburg, nurnberger, capital, hellersdorf, recklinghausen, meinfeldkirch, refugee home
17	News	today today the day, afternoon, everyday, afternoon, Saturday morning, daily, good morning, church day, morning, if possible, everyday, morning, enable, the day after tomorrow, state media, Friday, Monday, hour day, fridayforfuture, happybirthday, matchday, noon, Fridays, Valentine's Day, impossible

Notes: This table lists the most important topic words for the topics generated by the top2vec topic model (Angelov, 2020).

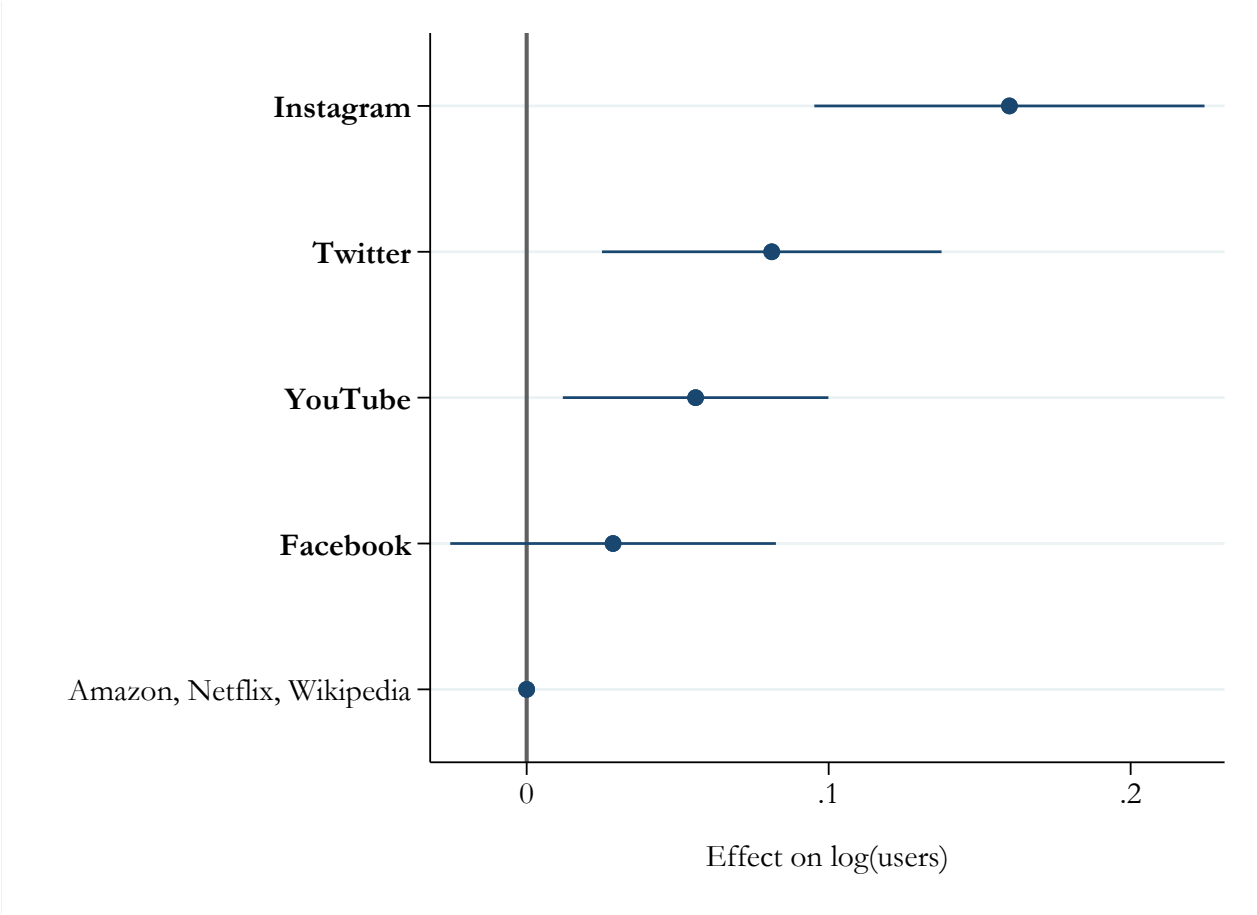
Table C.10: Overview Topic Model: Rising Topics

Topic Nr.	Topic Label	Topic Words
3	Germany	german, german turks, germans, germany, german, german, german, german, germany, deutschlandfunk, deutschebahn, germanywide, tedesco, germany, germania, deutschlan, deutschl, nazi march, deutsch, deutschla, niemehrcsu, east german, deutschebank, east german, nazis, east germany, north germany, niemehrcdu, berlin election, nazisraus, berlinale, berlin, west german, nazi, neo-nazis, south german, berlin, austria, neo-nazi, holocaust, hollande, berlindirekt, deutschrap, niemehrspd, wurttemberg, nationalists, ltwbayern, nationalism, migrants, amsterdam
20	Neo-nazis	nazi march, neo-nazis, nazis, nazisraus, neo-nazi, nazi, holocaust, fascist, anti-fascists, right-wing extremist, anti-fascism, fascist, anti-semitism, anti-semitic, anti-fascist, anti-semitic, anti-semitic, nationalists, anti-semitic, anti-semitic, german turks, extremists, nationalism, fascists, anti-semitic, deutschebahn, extremism, north germany, deutschlandfunk, fascism, neoliberalism, german, mass murderer, german, german, germany, germany, civil war, left-wing fascists, communists, german, german, neoliberal, socialist, socialist, german, berliner, berlinale, right-wing radical, jihadists
18	Feminism	feminists, feminists, feminist, feminist, feminism, feminist, feminist, sexism, sexist, sexist, women, female, patriarchy, female, ladies, gender, woman, female, muller, genders, gender, international women's day, politicians, lady, female, candidates, participants, cleaning lady, hannelore, ladies, masculinity, masculinity, comrades, teachers, girl, journalists, lesbian, sexual
45	Anti-semitism	anti-Semitism, anti-Semites, anti-Semitic, anti-Semitic, anti-Semitic, anti-Semitic, anti-Semitic, Jews, Jew-hatred, Jewish, Jewish, Israeli, Israeli, Israel, Jewish, Israelis, Jewish, Zionists, Israel, Jewish, Jew, holocaust, synagogues, neo-Nazis, synagogue, nazi march, neo-nazi, judaism, anti-fascists, judith, nazis, netanyahu, nazisraus, anti-fascism, jerusalem, jihadists, right-wing extremist, nazi, anti-fascist, dusseldorf, duesseldorf, hellersdorf, dusseldorfer, islamophobia, extremists, extremism, fascist, zehlendorf, fascist, islamists
7	Politics (Far-right/left)	people's party, left-wing party, anti-democrats, the party, people's parties, democrats, republicans, pirate party, democrat, right-wing extremist, old parties, fascist, anti-fascists, social democrat, state party conference, liberals, protest voters, liberal, federal party conference, workers' party, fascist, liberals, fascists, anti-fascist, women's suffrage, democratic, left-wing fascists, democratic, anti-fascism, parties, right-wing radical, right-wing radical, democratic, party, democratic, right-wing radical, liberal, extremists, liberalism, conservative, democratic, undemocratic, conservatives, political, parliament, political, party donations, parliamentary election, fascism, demonstrators
157	Local Events	Clausnitz, Mecklenburg, Chemnitz, Copenhagen, Heidenheim, Hoffenheim, Meinfeldkirch, Recklinghausen, Connewitz, Hellersdorf, Tecklenburg, Oberhausen, Hildesheim, Reinickendorf, Weinheim, Dusseldorf, capital, Duesseldorf, Naziaufmarsch, neo -Nazis, Holocaust, Russelladt, Ingolstadt, Dusseldorfer,, mulheim, holstein, berlin, zehlendorf, berlin election, berlinale, neustadt, sweden, magnitz, neo-nazi, switzerland, darmstadt, anti-semitism, stockholm, alexanderplatz, kimmich, stauffenberg, schanzenviertel, swedish, nordstadt, hessen election, heidelberg, wikileaks, nazis, demonstrators
72	Christmas	christmas time, christmas party, christmas festival, christmas, christmas, christmas, christmas tree, christmas money, santa claus, christmas market, christmas eve, new year's reception, holidays, holidays, holiday, natalie, new year's cleaning, new year's day, celebrates, valentine's day, christchurch, christ child, snowing, halloween, christopher, winter break, snowden, christiane, christian, christian, celebrate, christian, winter, christine, christianity, christ, christin, kristina, christian, christoph, winter time, january, christ, christina, gifts, winterthur, christian, celebrated, november, february
37	Politics (Conservatism/Liberalism)	people's party, the party, left party, people's parties, parliament, pkk, parliaments, parliaments, social democrat, republican, pirate party, old parties, parliamentary election, parliamentarians, socialist, socialists, socialism, socialist, democrats, politician, anti-democrats, politicians, local elections, goes elections, federal minister, democrat, new elections, berlin election, workers' party, liberal, politician, liberals, political, politician, political, conservative, politics, political, candidate for chancellor, conservatives, political, neoliberalism, politicians, political, political, protest voter, foreign minister, liberal, party donations
38	Justice System	legal, legal, constitutional state, legal, legal, legal, constitutional state, legal situation, legal, lawyer, constitutional state, legal, international law, legislation, law, unlawful, legal, basic law, legal, basic law, lawyer, legal system, lawyers, laws, legal, legislator, police law, court, criminal law, public prosecutor, prevention, bill, betting, court, law, lawyer, prosecutors, legal, legal committee, right-wing, law, criminal, asylum law, legalization, sued, arbitrator, just, civil rights, court of auditors, justice
22	Journalism	journalist, journalists, journalist, journalism, clipmedianews, journalist, journalists, media report, media, newsticker, daily newspaper, reporter, newsbasel, reporter, anonymousnews, newsflash, press conference, newsletter, news, media library, newsroom, srfnews, propaganda, freedom of the press, pers, press mirror, journal, press spokesman, koran, medial, medial, reports, social media, press, press club, mediale, magazine, press, reportage, news, media, printed, publication, publish, wikileaks, multimedia, liegenpresse, magazine, print out, tv

Notes: This table lists the most important topic words for the topics generated by the top2vec topic model (Angelov, 2020).

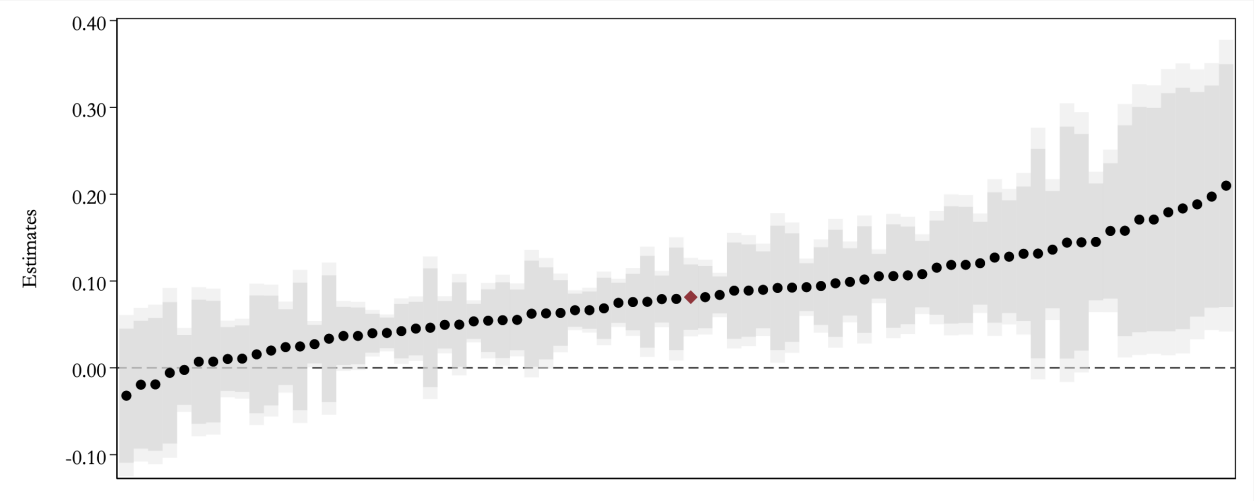
C.4. Additional Results for Platform Usage

Figure C.4: The Effect of the NetzDG on Platform Usage by Platform



Notes: This figure plots coefficients from a version of Equation (2) which replaces the dummy for treated platforms with dummies for individual platforms. The dependent variable is the log number of users. The omitted category is the set of untreated platforms. The whiskers indicate 95% confidence intervals based on standard errors clustered by country.

Figure C.5: The Effect of the NetzDG on Platform Usage, Specification Curve



Notes: This figure plots coefficients from a version of Equation (2) which changes the set of treated and control platforms. We consider all possible combinations with at least one control platform and two treated platforms. The dependent variable is the log number of users. The gray bars denote 95% confidence intervals based on standard errors clustered by country. The red rhomboid denotes the specification reported in the paper.

D Additional Results on Offline Effects

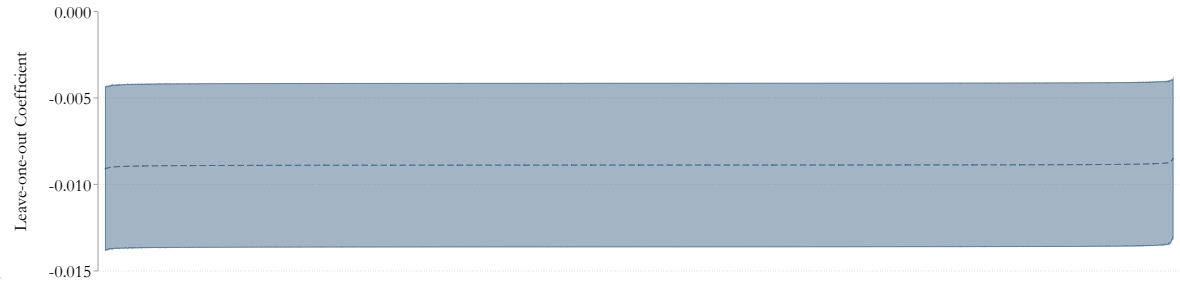
D.1. Additional Results for Hate Crimes

Table D.1: Robustness: Type of Hate Crime Incident

	Dep. var.: Type of Anti-refuge Hate Crime					
	All	Arson	Assault	Demonstration	Other	Suspect. Cases
	(1)	(2)	(3)	(4)	(5)	(6)
AfD Facebook users p.c. (std) × Post	-0.009***	-0.000	-0.003**	-0.001**	-0.008***	-0.001**
	(0.002)	(0.000)	(0.001)	(0.000)	(0.002)	(0.000)
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes
Year-Quarter FE	Yes	Yes	Yes	Yes	Yes	Yes
Ln(Pop/) × Post	Yes	Yes	Yes	Yes	Yes	Yes
AfD vote share × Post	Yes	Yes	Yes	Yes	Yes	Yes
Election Controls × Post	Yes	Yes	Yes	Yes	Yes	Yes
Facebook users p.c × Post	Yes	Yes	Yes	Yes	Yes	Yes
Broadband internet × Post	Yes	Yes	Yes	Yes	Yes	Yes
Observations	71,008	71,008	71,008	71,008	71,008	71,008
Pre-Period Mean of DV	0.12	0.00	0.02	0.00	0.10	0.01
R^2	0.44	0.09	0.38	0.15	0.40	0.16

Notes: This table presents the results of estimating municipality-quarter-level regressions as in Equation (3) where the dependent variable is the inverse hyperbolic sine of the number of anti-refugee hate crimes of a specific type (indicated in the top row). *AfD Facebook users p.c. (std)* is the number of AfD Facebook followers per capita, standardized to have a mean of 0 and a standard deviation of 1 to ease interpretation. All regressions include municipality and quarter fixed effects, as well as controls for the logarithm of population, the AfD vote share, Facebook users per capita, and broadband internet access, all interacted with *Post*. Robust standard errors in parentheses are clustered by county. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Figure D.1: Leave-one-out Estimates



Notes: This figures shows the estimates of a leave-one-out exercise, where we estimate Equation (1) omitting one municipality at a time. The figure plots a total of 4,466 estimates sorted by size. The dashed line are the point estimate and the shading indicates 95% confidence intervals.

Table D.2: Robustness: Specification

	Dep. var.: Anti-Refugee Hate Crime					
	Asinh	Count	Ln(p.c.)	Asinh	Count	Ln(p.c.)
	(1)	(2)	(3)	(4)	(5)	(6)
AfD Facebook users p.c. (std) × Post	-0.009***	-0.022***	-0.007***			
	(0.002)	(0.005)	(0.002)			
High AfD Usage × Post				-0.023***	-0.065***	-0.018***
				(0.007)	(0.018)	(0.005)
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes
Year-Quarter FE	Yes	Yes	Yes	Yes	Yes	Yes
Ln(Pop/) × Post	Yes	Yes	Yes	Yes	Yes	Yes
AfD vote share × Post	Yes	Yes	Yes	Yes	Yes	Yes
Election Controls × Post	Yes	Yes	Yes	Yes	Yes	Yes
Facebook users p.c × Post	Yes	Yes	Yes	Yes	Yes	Yes
Broadband internet × Post	Yes	Yes	Yes	Yes	Yes	Yes
Observations	71,008	71,008	71,008	71,008	71,008	71,008
Pre-Period Mean of DV	0.12	0.19	-9.06	0.12	0.19	-9.06
R^2	0.44	0.63	0.95	0.44	0.63	0.95

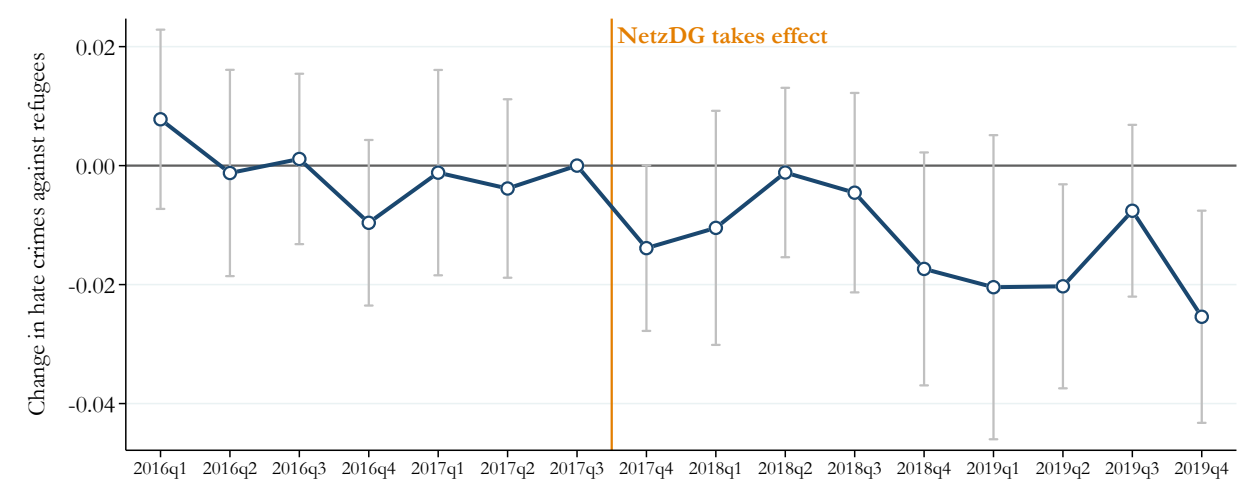
Notes: This table presents the results of estimating municipality-quarter-level regressions as in Equation (3) where the dependent variable is the transformation of anti-refugee hate crimes indicated at the top of the table. *AfD Facebook users p.c. (std)* is the number of AfD Facebook followers per capita, standardized to have a mean of 0 and a standard deviation of 1 to ease interpretation. *High AfD Usage* is an indicator equal to 1 for municipalities with an above-median number of AfD Facebook followers per capita. All regressions include municipality and quarter fixed effects, and controls for the logarithm of population, the AfD vote share, Facebook users per capita, and broadband internet access, all interacted with *Post*. Robust standard errors in parentheses are clustered by county. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table D.3: Robustness: Standard Errors

	Standard Errors Clustered by:			
	County	County & Quarter	Municipality	Municipality & Quarter
	(1)	(2)	(3)	(4)
AfD Facebook users p.c. (std) × Post	-0.009***	-0.009***	-0.009***	-0.009***
	(0.002)	(0.002)	(0.002)	(0.002)
Municipality FE	Yes	Yes	Yes	Yes
Year-Quarter FE	Yes	Yes	Yes	Yes
Ln(Pop/) × Post	Yes	Yes	Yes	Yes
AfD vote share × Post	Yes	Yes	Yes	Yes
Election Controls × Post	Yes	Yes	Yes	Yes
Facebook users p.c × Post	Yes	Yes	Yes	Yes
Broadband internet × Post	Yes	Yes	Yes	Yes
Observations	71,008	71,008	71,008	71,008
Pre-Period Mean of DV	0.12	0.12	0.12	0.12
R^2	0.44	0.44	0.44	0.44

Notes: This table presents the results of estimating municipality-quarter-level regressions as in Equation (3) where the dependent variable is the inverse hyperbolic sine of the number of anti-refugee hate crimes. *AfD Facebook users p.c. (std)* is the number of AfD Facebook followers per capita, standardized to have a mean of 0 and a standard deviation of 1 to ease interpretation. All regressions include municipality and quarter fixed effects, as well as controls for the logarithm of population, the AfD vote share, Facebook users per capita, and broadband internet access, all interacted with *Post*. Robust standard errors in parentheses are clustered at the level indicated at the top of the table. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Figure D.2: Event Study Hate Crime (Twitter Exposure)



Notes: This figure plots the coefficients from running an event study version of regression Equation (3). The dependent variable is the inverse hyperbolic sine of the number of anti-refugee incidents. Exposure is measured based on the number of AfD Twitter followers per capita in each municipality. The omitted category is the 3rd quarter of 2017, the quarter before the passing of the NetzDG (indicated with the vertical line). The whiskers indicate 95% confidence intervals based on standard errors clustered by county.

Table D.4: Robustness: Social Media Exposure measured with Twitter

	Dep. var.: Asinh(Anti-Refugee Hate Crimes)				
	(1)	(2)	(3)	(4)	(5)
AfD Twitter Followers p.c. (std) × Post	-0.012**	-0.010**	-0.011**	-0.010**	-0.011**
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
AfD vote share (std) × Post		0.033***	0.032***	0.034***	0.029**
		(0.012)	(0.012)	(0.012)	(0.012)
Facebook users p.c (std) × Post			0.002	0.002	0.003
			(0.002)	(0.002)	(0.002)
Broadband internet (std) × Post				0.004	0.000
				(0.004)	(0.004)
Municipality FE	Yes	Yes	Yes	Yes	Yes
Year-Quarter FE	Yes	Yes	Yes	Yes	Yes
Ln(Pop.) × Post	Yes	Yes	Yes	Yes	Yes
Election Controls × Post		Yes	Yes	Yes	Yes
All Controls (19) × Post					Yes
Observations	71,456	71,008	71,008	71,008	68,736
Pre-Period Mean of DV	0.12	0.12	0.12	0.12	0.12
R^2	0.44	0.44	0.44	0.44	0.45

Notes: This table presents the results of estimating Equation (3), where the dependent variable is the inverse hyperbolic sine of the number of anti-refugee hate crimes in a municipality in a given quarter. *AfD Twitter Followers p.c. (std)* is the number of AfD Twitter Followers per capita, standardized to have a mean of 0 and a standard deviation of 1 to ease interpretation. All regressions include municipality and quarter fixed effects as well as a control for the natural logarithm of population, interacted with *Post*. See text for a detailed description of the additional control variables. Robust standard errors in parentheses are clustered by county. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table D.5: Mechanism: Heterogeneity by Population Size and Density

	Dep. var.: Asinh(Anti-Refugee Hate Crimes)					
	All Incidents		Single Perp.		Multiple Perp.	
	(1)	(2)	(3)	(4)	(5)	(6)
AfD Facebook users p.c. (std) × Post	0.001	-0.001	-0.000	0.000	0.001	0.000
	(0.002)	(0.002)	(0.000)	(0.000)	(0.000)	(0.000)
AfD Facebook users p.c. (std) × Post × High Population	-0.036***		-0.004***		-0.012***	
	(0.007)		(0.001)		(0.003)	
AfD Facebook users p.c. (std) × Post × High Population Density		-0.024***		-0.004***		-0.009***
		(0.005)		(0.001)		(0.003)
Municipality FE	Yes	Yes	Yes	Yes	Yes	
Year-Quarter FE	Yes	Yes	Yes	Yes	Yes	
Ln(Pop.) × Post	Yes	Yes	Yes	Yes	Yes	Yes
Observations	71,456	71,456	71,456	71,456	71,456	71,456
Pre-Period Mean of DV	0.118	0.118	0.005	0.005	0.012	0.012
R^2	0.44	0.44	0.14	0.14	0.20	0.20

Notes: This table presents the results of estimating Equation (3), where the dependent variable is the inverse hyperbolic sine of the number of anti-refugee hate crimes in a municipality in a given quarter. We report the estimates for all anti-refugee incidents in columns (1-2) and restrict the sample to hate crimes committed by single perpetrators in columns (3-4) and multiple perpetrators in columns (5-6). *AfD Facebook users p.c. (std)* is the number of AfD Facebook followers per capita, standardized to have a mean of 0 and a standard deviation of 1 to ease interpretation. We additionally include interactions with indicators for high population or high population density municipalities. We additional control for the two-way interactions between either the high population or high population density indicator with *Post*. All regressions include municipality and quarter fixed effects as well as a control for the natural logarithm of population, interacted with *Post*. See text for a detailed description of the additional control variables. Robust standard errors in parentheses are clustered by county. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

D.2. Additional Results on Refugee Attitudes

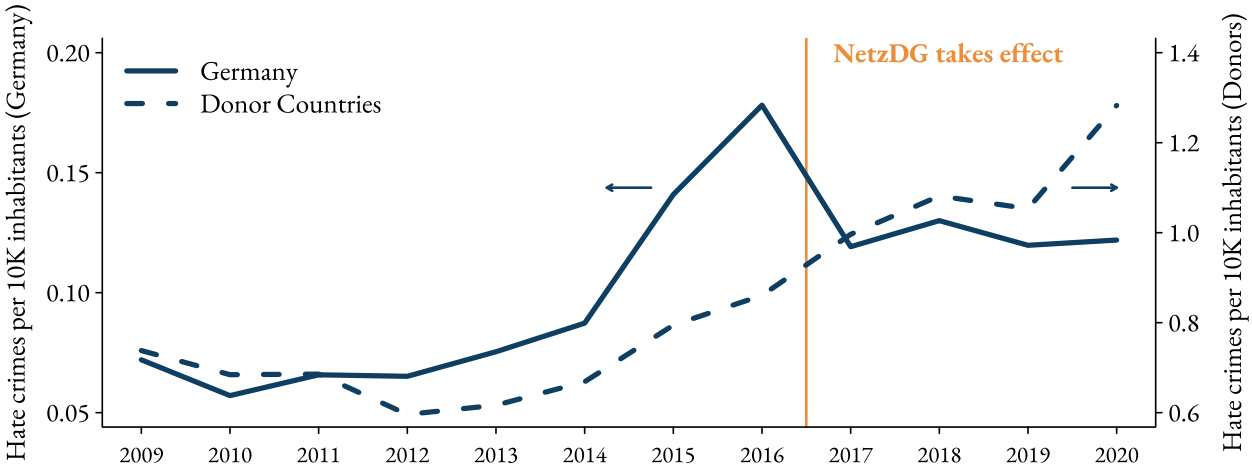
Table D.6: Changes in Action Towards Refugees

	Dep. var.: Helped Refugees by ...						
		Donated		Volunteered		Demonstrated	
	Index	Last Year	Future	Last Year	Future	Last Year	Future
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel (a): All Respondents							
Social Media User × Post	-0.005	-0.004	-0.008	-0.006	-0.011*	-0.006	0.005
	(0.004)	(0.008)	(0.008)	(0.005)	(0.006)	(0.004)	(0.006)
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Survey Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	36,786	36,560	35,586	36,388	35,404	36,336	35,360
Pre-Period Mean of DV	0.17	0.31	0.36	0.07	0.13	0.05	0.09
R^2	0.78	0.73	0.73	0.71	0.68	0.69	0.69
Panel (b): AfD Voters							
Social Media User × Post	-0.002	0.018	0.002	0.010	-0.005	-0.014	-0.013
	(0.011)	(0.020)	(0.021)	(0.014)	(0.015)	(0.015)	(0.024)
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Survey Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2,554	2,544	2,506	2,544	2,502	2,534	2,488
Pre-Period Mean of DV	0.07	0.10	0.10	0.02	0.03	0.05	0.10
R^2	0.68	0.68	0.72	0.69	0.62	0.66	0.62

Notes: This table presents the results of estimating Equation (4), where the dependent variables are different measures for positive actions towards refugees. *Social Media Users* is an indicator for respondents who use social media at least once a week. All regressions include individual and survey year fixed effects as well as a control for the respondent's gender and age, interacted with *Post*. See the text for a detailed description of the variables. Robust standard errors in parentheses are clustered by individual. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

D.3. Additional Synthetic Control Results

Figure D.3: Evolution of Hate Crimes in Germany vs. Donor Countries



Notes: This figure compares hate crimes per 10K inhabitants in Germany vs. the unweighted average in the donor countries in 2009-2020.

Table D.7: Country Weights in the Synthetic Germany

Country	Weight
Austria	0.09
Belgium	0.01
Bosnia and Herzegovina	0
Bulgaria	0
Croatia	0
Cyprus	0
Czech Republic	0
Denmark	0.01
Finland	0
Italy	0.1
Lithuania	0.55
Moldova	0.07
Poland	0.12
Portugal	0
Slovakia	0
Spain	0
Switzerland	0
Turkey	0.03
UK	0
Ukraine	0
US	0

Notes: This table presents the country weights used to generate the synthetic version of Germany for the synthetic control estimates.

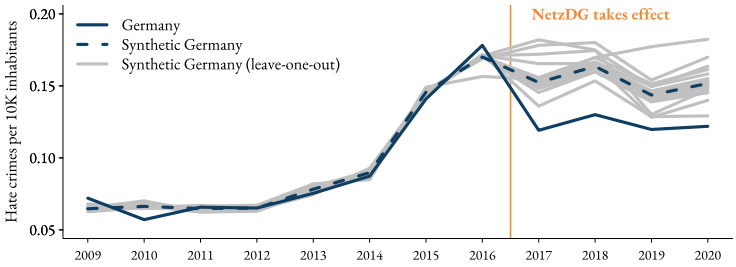
Table D.8: Hate Crimes Predictor Means

Variable	Germany		Donors	OECD	OSCE
	Real	Synthetic			
Hate crimes per 10K inhabitants 2009	0.07	0.06	0.74	1.02	0.71
Hate crimes per 10K inhabitants 2010	0.06	0.07	0.68	0.93	0.66
Hate crimes per 10K inhabitants 2011	0.07	0.06	0.69	0.92	0.66
Hate crimes per 10K inhabitants 2012	0.07	0.07	0.6	0.75	0.57
Hate crimes per 10K inhabitants 2013	0.08	0.08	0.62	0.73	0.59
Hate crimes per 10K inhabitants 2014	0.09	0.09	0.67	0.83	0.64
Hate crimes per 10K inhabitants 2015	0.14	0.15	0.79	1.03	0.76
Hate crimes per 10K inhabitants 2016	0.18	0.17	0.86	1.18	0.83
Measure change 2009-2016	0	0.11	0.11	0.11	0.11

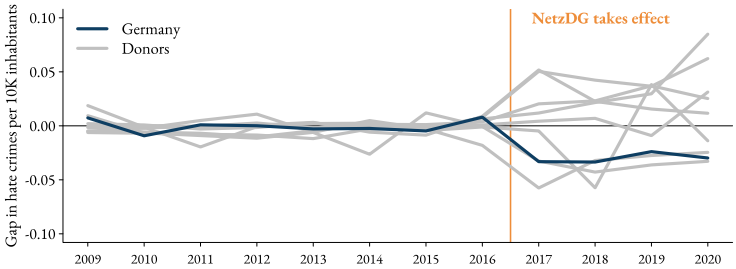
Notes: This table presents the means of the predictor variables for Germany and the synthetic Germany, as well as the simple mean among the donor, OECD, and OSCE countries.

Figure D.4: Leave-One-Out and In-Space Placebos

(a) Germany vs. Synthetic Control

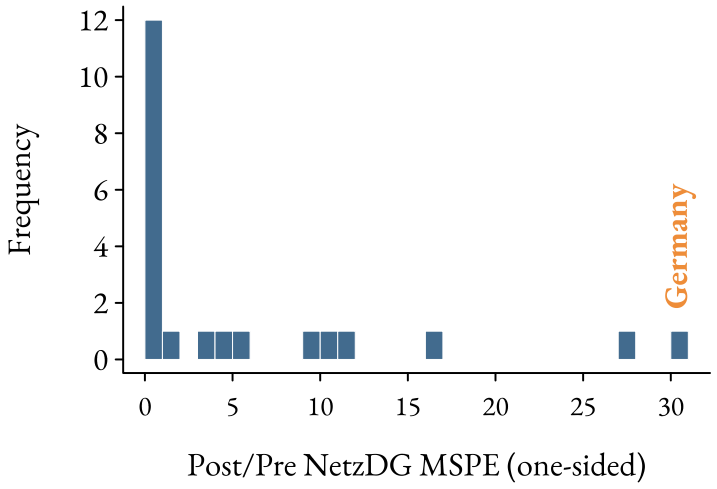


(b) Gaps Between Observed and Synthetic Hate Crimes



Notes: Panel (a) compares hate crimes per 10K inhabitants in Germany vs. the synthetic Germany and a synthetic Germany built by dropping each of the donor countries. Panel (b) shows the gaps between observed and synthetic values for Germany and each of the donor countries acting as a "placebo" treated country. As in Abadie et al. (2010), we drop countries with a pre-NetzDG MSPE higher than 5 times the one of Germany to improve the visibility of the graph.

Figure D.5: Mean Squared Prediction Error Ratios (One-Sided)



Notes: This graph plots the histogram of the ratio between the MSPE post-NetzDG and the MSPE pre-NetzDG. One-sided MSPE are calculated as in Abadie (2021).

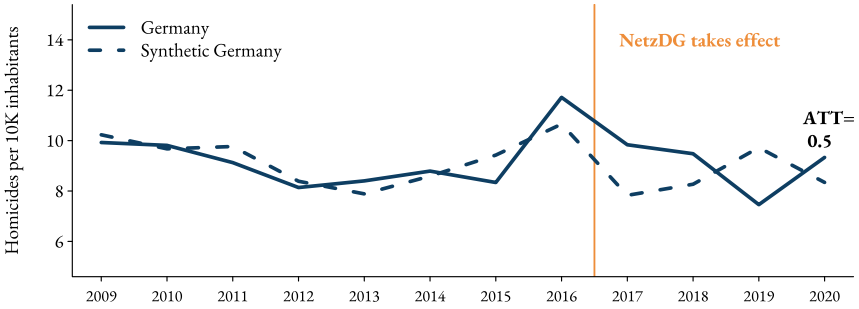
Table D.9: Robustness to Alternative Specifications

Specification	ATT	p-value (one-sided)	p-value (two-sided)	Donors	Pre-NetzDG RMSPE
Baseline	**-0.03**	**0.045**	**0.227**	**21**	**0.005**
Alternative interpolation					
Interpolation dummy	-0.03	0.045	0.182	21	0.005
No interpolation	-0.048	0.167	0.167	11	0.01
Alternative outcomes					
Log	-0.097	0.05	0.1	19	0.005
Levels	-0.047	0.136	0.318	21	0.012
Alternative periods					
Period 2009-2019	-0.051	0.042	0.042	23	0.005
Period 2009-2021	-0.086	0.056	0.111	17	0.007
Alternative donors					
Leave-one-out (max ATT)	-0.014	0.19	0.524	20	0.006
Leave-one-out (min ATT)	-0.048	0.048	0.238	20	0.009
OECD	-0.067	0.067	0.133	14	0.007
No Lithuania	-0.036	0.048	0.143	20	0.006

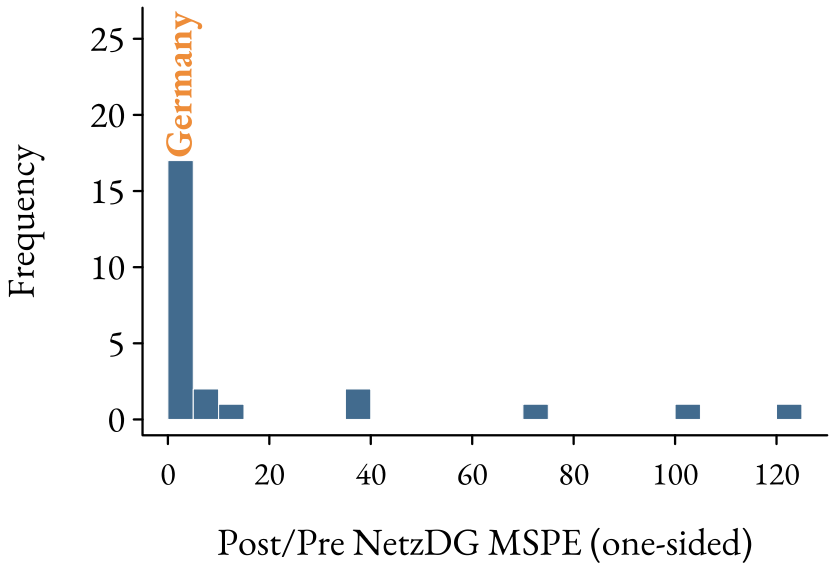
Notes: This table presents estimates of the average treatment effect post-NetzDG, its one- and two-sided p-values, the number of donors and the pre-NetzDG root mean squared prediction error. Note that the ATT and the RMSPE are expressed in hate crimes per 10K inhabitants, to facilitate comparison between specifications. Inference is based on the permutation method of Abadie et al. (2010); see Abadie (2021) for how to compute one-sided p-values. "Interpolation dummy" adds as predictor the pre-NetzDG average of a dummy indicating observations that were linearly interpolated. "No interpolation" keeps only countries without missing values during the period of study.

Figure D.6: Placebo Outcome: Homicides

(a) Synthetic Control Estimates



(b) Mean Squared Prediction Error Ratios (One-Sided)



Notes: Panel (a) shows synthetic control estimates. The figure presents the evolution of homicides per 10,000 inhabitants in Germany and the synthetic Germany. The synthetic control uses all lagged outcomes as predictors, as well as the average of a dummy variable indicating whether there were measurement changes in the hate-crime series in the pre-period. Panel (b) plots the histogram of the ratio between the MSPE post-NetzDG and the MSPE pre-NetzDG. One-sided MSPE are calculated as in Abadie (2021).