The Effect of Content Moderation on Online and Offline Hate: Evidence from Germany's NetzDG*

Rafael Jiménez Durán† Karsten Müller‡ Carlo Schwarz§

February 28, 2023

Abstract

Social media companies are under scrutiny for the prevalence of hateful content on their platforms, but there is scarce empirical evidence of the consequences of regulating such content. We study this question with a particular focus on anti-refugee hate crime in the context of the "Network Enforcement Act" (NetzDG) in Germany, which mandates major social media companies to remove hateful posts within 24 hours. Using a difference-in-differences strategy, we find that the law was associated with a 4% reduction in the toxicity of refugee-related tweets by far-right social media users. Further, we show that the NetzDG reduced anti-refugee hate crimes in municipalities with more far-right social media users. The estimates suggest that the NetzDG induced a -0.9 percentage point reduction reduction in anti-refugee incidents for every standard deviation of far-right social media usage. These findings are also confirmed by a synthetic control estimate. Together, these results suggest that online content moderation can curb online hate speech and offline violence.

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

*We are grateful to Leonardo Bursztyn, Fabrizio Germano, Sophie Hatte, Ro'ee Levy, Sulin Sardoschau, David Yang, Noam Yuchtman, Ekaterina Zhuravskaya and seminar participants at Bocconi University, the Winter Meeting Econometric Society, University of Cologne, and Toulouse School of Economics for their helpful suggestions. Any errors are ours.

†Social Science Research Council, Stigler Center, rafaeljjd@uchicago.edu.
‡National University of Singapore, Department of Finance, 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, such as Facebook and Twitter, 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 on the impact of social media on attacks against ethnic and religious minorities suggests that there are indeed grounds for these concerns (see Müller and Schwarz, 2021, 2022b; Bursztyn et al., 2019).

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 Twitter since 2015 (Gillespie, 2018; Twitter, 2015). But these content moderation attempts remain controversial: some people object that platforms are not moderating enough, while others are concerned about online censorship. Before evaluating whether such policies are socially desirable, however, it is crucial to understand whether they can effectively reduce online hate and its violent offline consequences.

This paper sheds light on this question by focusing on the first legal change explicitly aimed at increasing the moderation efforts of social media platforms: 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 during the 2015-2016 refugee crisis. The law marks a unique and unprecedented legal change that introduced large penalties for social media platforms, of up to €50 million, for failing to promptly remove hateful content.[1] As such, the law drastically changed social media providers' incentives to remove hateful content, and has been called a "key test for combatting hate speech on the internet" (Echikson and Knodt, 2022).

In this paper, we investigate whether increased content moderation efforts induced by the NetzDG indeed decreased online and offline hatred targeting refugees. We put a particular focus anti-refugee online content and anti-refugee hate crime as these have been linked by previous research (see Müller and Schwarz, 2021). Therefore, we analyze the Twitter and Facebook accounts of followers of the Alternative for Germany

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

("Alternative für Deutschland", henceforth AfD). The AfD, at the time the NetzDG became effective, was the third-largest party in the German parliament, having risen on a platform of far-right anti-immigrant rhetoric, with a particular focus on refugees. Importantly, the AfD also had (and still has) the largest Facebook following of any German party.

In the first part of our empirical analysis we provide evidence that the NetzDG was indeed followed by a decrease in the toxicity of social media posts, as measured by Google's Perspective API, an algorithm commonly used in industry applications and as a benchmark in academic studies. In a difference-in-differences analysis with Twitter data, we compare the content produced by "toxic users" and "non-toxic users" before and after the NetzDG was implemented. Intuitively, users who posted more toxic tweets before the law was passed should be more exposed to increases in online content moderation. To obtain a comparable set of users, we sample accounts who post the word "refugee" in German. Among these, we consider two criteria to select "toxic" users: those who follow the AfD account on Twitter, and those whose pre-NetzDG toxicity is above a certain percentile.

We find a significant decrease in the toxicity of both refugee-related tweets and overall tweets after the law was passed irrespective of how we define "toxic" users. Compared to other users, we observe a drop in online toxicity of 32% for those in the top decile of toxicity pre-NetzDG and a drop of around 4% for AfD followers. These findings are in line with the predictions from a simple theoretical framework that models content moderation as a quality decision for platforms and the NetzDG as a tax on unmoderated content.

In the second part of our empirical analysis, we investigate the effects of the NetzDG on hate crimes against refugees, exploiting municipality-level differences in the exposure to far-right social media content. If the NetzDG limited online hate speech, as we have shown in first part, 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 two-way fixed effects regressions, we find that the introduction of the NetzDG led to a reduction of anti-refugee incidents in municipalities with many AfD Facebook followers. The estimates suggest that municipalities with one standard deviation higher number of AfD followers per capita saw a -0.9 percentage point reduction in the number of anti-refugee incidents.

We also investigate the intensive margin of far-right Facebook usage. Specifically, we find a stronger reduction of anti-refugee hate crimes, over and above what is predicted

by the number of AfD followers, depending on the frequency with which users interact with the AfD's Facebook page (as measured by posts, likes, comments, or shares). For example, municipalities with one standard deviation higher number of posts per AfD follower experience a further -0.5 percentage point reduction in the number of anti-refugee hate crimes after the NetzDG.

The underlying identification assumption of our 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. While this assumption is inherently untestable, 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 other municipality characteristics and a battery of robustness checks. For example, our estimates are not driven by differences in local social media or internet penetration, the number of refugees, nor by strong support for the AfD in the 2017 federal election. If anything, the coefficients for these variables are often positive, although they are mostly statistically insignificant. The results are also unlikely to be driven by the end of the refugee crisis itself. Furthermore, the main results remain unchanged if we consider Twitter-based exposure measures, alternative variable transformations, standard errors, more restrictive fixed effects, and sub-samples for our analysis.

To get a sense of the aggregate effect of the policy, we additionally build a synthetic control for Germany using 2009-2020 data from 22 donor countries, following the methodology introduced by Abadie and Gardeazabal (2003) and Abadie et al. (2010). Using the full path of pre-intervention outcomes 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.

Related Literature Our paper contributes to three strands of the literature. First, there is a fast-growing literature on the real-life outcomes of social media. Existing work has investigated the impact of social media 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), and voting (Bond et al., 2012; Jones et al., 2017; Fujiwara et al., 2021). Zhuravskaya et al. (2020) review the recent literature on the political effects of social media. Most closely related are three papers that provide evidence for the impact of social media on hate

crimes (Müller and Schwarz, 2021, 2022b; Bursztyn et al., 2019). Despite this existing work, we know little about how to effectively curb the adverse real-world effects of hateful messaging on social media. To the best of our knowledge, our paper is the first to show that content moderation policies can have real-life effects.

Second, we contribute to a nascent literature that studies platform decisions and content moderation strategies (Acemoglu et al., 2021; Liu et al., 2021; Madio and Quinn, 2021). Jiménez Durán (2022) finds that changing beliefs about content moderation has an insignificant effect on consumer surplus. This finding 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. Müller and Schwarz (2022a) study the aftermath of Donald Trump's account deletion and document decreases in online toxicity but also platform engagement. Our first-stage findings are also in line with the work of Andres and Slivko (2021), who provide suggestive evidence that the toxicity of far right-wing German Twitter users decreased after the NetzDG relative to a set of Austrian Twitter users.

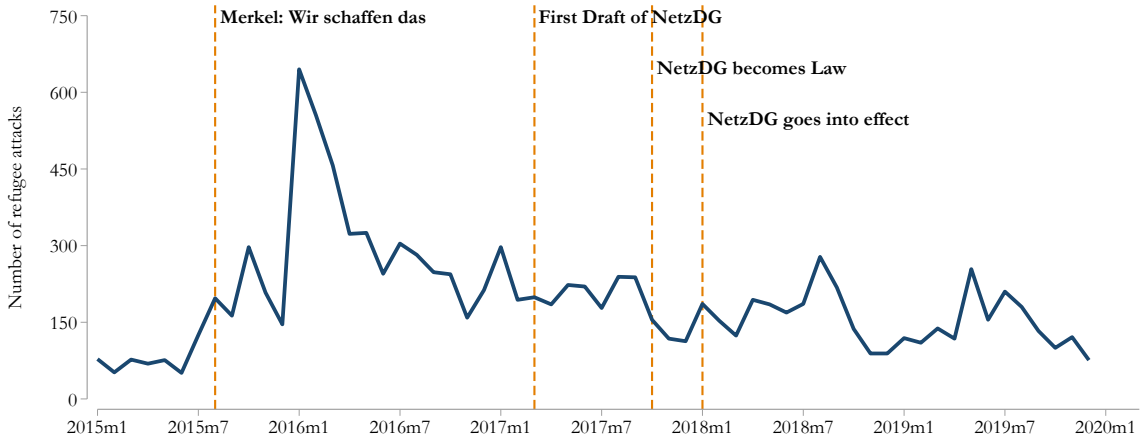
Lastly, we speak to the literature on the effects of traditional media and violence. Research by Yanagizawa-Drott (2014), DellaVigna et al. (2014), and Adena et al. (2015), for example, suggests that nationalist propaganda on the radio can increase the prevalence violence against minorities. In other 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. Relative to this literature, our paper not only studies an environment that is far less regulated than traditional media, but also a media platform that allows the active participation of users.

2 Background

In August 2015, Chancellor Angela Merkel declared that Germany would welcome a large number of refugees of 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 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.

The large inflow of asylum seekers into Germany was also accompanied by a flare-up in the number of anti-refugee incidents in Germany. The non-profit organization "Amadeu Antonio Stiftung" recorded more than 11,620 hate crimes targeting refugees in Germany between 2015 and 2020, visualized in Figure 1. These 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).

Figure 1: Time Series of Attacks on Refugees in Germany



Notes: This time series plot shows the monthly number of refugee attacks in Germany between 2015 and 2019. The dashed vertical lines mark the date of Merkel's "Wir schaffen das!" speech and important dates in the creation and approval of the NetzDG.

In previous research, Müller and Schwarz (2021) have shown that social media played a role in this wave of anti-refugee crime. The evidence suggests that far-right Facebook pages helped propagate anti-refugee sentiment, and the exposure to such online content motivated real-world anti-refugee incidents. The Facebook page of Alternative for Germany (AfD) became a key platform for the spread of anti-refugee content.

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 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.[2] 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 in September 2017. It became law in October 2017 and went into force on January 1st, 2018.

The NetzDG was "the first law that formalises the process for platform takedown obligations" (Kohl, 2022). 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, the 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 notice. To incentivize users to report hateful content, the NetzDG required platforms to add a dedicated button to report violations against the law. Appendix Figure A.2 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).

In Online Appendix A.1., we provide a theoretical framework that allows us to derive predictions about the impact of the NetzDG on the prevalence of hateful content. Within the framework, the NetzDG can be interpreted 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, where it 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. We confirm the predictions of this model in the first part of the paper.

In the next section, we describe our main data sources and the empirical strategy that will allow us to investigate the impact of the NetzDG on online hate speech and

[2]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.

offline hate crimes.

3 Data

We construct three separate data sets for our analysis. First, we build a database of refugee-related tweets that allows us to study the impact of the NetzDG on the toxicity of online content. Second, for our main analysis on the spillovers of online hate speech into real-life action, we construct a municipality-quarter panel of anti-refugee incidents. Third, for our synthetic control analysis, we build a cross-country panel of total hate crime. We describe the main data sources for each of the data sets in the following.

3.1 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 data set measuring online toxicity of refugee-related tweets. We focus on data from Twitter 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 was also one of ten platforms 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 increases in online hate speech that occured 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 toxicity of refugee-related German tweets. In total, this dataset contains 346,167 tweets. We also provide evidence for the effects of the NetzDG on the toxicity of the overall discourse on Twitter by collecting all other tweets of the users who were returned in the full-archive search of refugee-related content.

To identify the politcal leaning of users, we additionally scraped the Twitter follower lists of all major German parties. These lists allow us to identify which Twitter users are following the AfD's Twitter account.

We measure the toxicity of online content using Google's Perspective API (Wulczyn et al., 2017; Dixon et al., 2018). This API returns toxicity measures along the following six dimensions: toxicity, severe toxicity, identity attack, insult, profanity, and threat. Appendix Table A.3 contains summary statistics for our sample of refugee tweets. On

average, refugee-related tweets have a toxicity score equal to 0.41 and 3% of them had a toxicity score of at least 0.8, which is a commonly-used cutoff for classifying hate speech in the literature (ElSherief et al., 2018; Han and Tsvetkov, 2020; Vidgen et al., 2020).

As a benchmark, in a random sample of tweets in English, 5.6% of them had a toxicity score of at least 0.8 (Jiménez Durán, 2022)—the same prevalence we find in our data. 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 28% of tweets in the sample were posted by AfD followers and 49% of them were posted by users following at least one political party. Appendix Figure A.1 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.

3.2 Municipal Anti-Refugee Hate Crime Panel

Our main analysis is based on a panel data set 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).[3] Information on around three-quarters of these incidents comes from administrative police data reported based on parliamentary requests. All of the 10,081 anti-refugee crimes are classified into four groups. The most common cases are property damage to refugee homes (7,815 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 are able to link incidents to their corresponding municipality because they are geo-coded with exact longitude and latitude. We assign these incidents to municipalities using shape files provided by the ©GeoBasis-DE/BKG 2016 website.[4]

Most of the additional municipality-level variables ared based on the replication data from Müller and Schwarz (2021)[5]. We briefly describe each type of data we use below and refer the reader to Müller and Schwarz (2021) for additional details.

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

[4]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 there are smaller differences in the size and population of these administrative areas.

[5]The underlying reproduction file is available here.

The main measures of far-right Facebook usage from Müller and Schwarz (2022b) which we use in our analysis is 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. The motivation to use the AfD's Facebook page is that the AfD is a relatively new right-wing populist party whose Facebook page is arguably the key platform for anti-refugee content online 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 the data with information about the activity of each user. This allow us to construct the number of posts, likes, comments, and shares for each AfD user. [6]

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.

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

Finally, we also add municipality-level socio-economic controls 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 data on the share of the population that are immigrants and asylum seekers. Data on the availability of broadband internet 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

[6]The shares were not included in the replication file but stem from the same Facebook scraping.

Gesellschaft der deutschen Zeitungen (ZMG)" (Society for Market Research of German Newspapers), which we normalize using a municipality's population. Data on other types of crimes by county and year come from the Bundeskriminalamt (BKA)'s Police Crime Statistics.

We visualize the main variation in Figure 2. The map shows 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.1 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 are 10,080 anti-refugee incidents in our sample. 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.

3.3 Cross-Country Hate Crime Panel

We additionally construct a panel of hate crime incidents in different countries for the years 2009-2020, which will enable us to estimates results based on a synthetic Germany. The most comprehensive cross-country hate crime database 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 any data measurement changes over time.[7]

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 of violent hate crimes (which do no 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 attempt), Robberies Physical Injuries, and Violent Property damages. Lastly, we gathered population counts from the World Bank's World Development Indicators and refugee statistics from the United Nations Refugee Agency (UNHCR).

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

Table A.4 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.2. 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 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 data set contains 21 countries in addition to Germany. Figure A.7 shows the evolution of hate crimes in Germany and the raw mean of the donor countries. Unsurprisingly, we find that due to completely different pre-trends between countries traditional differences-in-differences analysis are not possible.

4 Empirical Strategy

Our empirical analysis proceeds in three steps. First, we provide evidence that the NetzDG reduced the toxicity of online content. Second, we study the relative effect of the policy on the frequency of anti-refugee hate crimes with a between-municipalities comparison Third, we estimate the aggregate effect of the policy by constructing a synthetic control for German hate crimes.

Online Effects. To investigate the effect of the NetzDG on hateful online content, we estimate a difference-in-differences regression of the following form:

$$Toxicity_{iut} = \theta \cdot Toxic\,User_u \times Post\;NetzDG_t + \phi_u + \mu_t + \psi_{iut}, \tag{1}$$

where $Toxicity_{iut}$ denotes the toxicity score of tweet i posted by user u on day t, based on the coding from the Google Perspective API. The main independent variable is the interaction between an indicator variable for highly toxic users ($Toxic\,User_u$) and the post-NetzDG dummy ($Post\;NetzDG_t$). $Post\;NetzDG_t$ is equal to 1 starting in the fourth quarter of 2017 (October 1, 2017), when NetzDG took effect. We show the results for two versions of $Toxic\,User_u$: one version that defines exposed users as those that sent particularly toxic content before the NetzDG, and one version based on the AfD's Twitter followers. At baseline, highly toxic users are defined as users above the 75th percentile of the pre-period toxicity distribution. In Appendix Table A.5, we show that our results hold irrespective of the chosen cutoff.

Our strategy compares the change in toxicity of refugee-related tweets posted by users producing particularly toxic content to other Twitter users, before and after the implementation of the NetzDG. Intuitively, we expect to see a decrease in the average toxicity of refugee-related tweets posted by more "exposed" users relative to others. Technically, the NetzDG also applied to content that was posted before the law was passed, but since the NetzDG relied on users flagging hateful 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 content posted *after* it went into force. Note that any content moderation of social content that was posted before the NetzDG would only bias our results towards 0 and our estimates can therefore be interpreted as a lower bound.

The Netz could decrease online toxicity either due to the removal of toxic posts or by deterring users from posting toxic content. The NetzDG could deter users from posting toxic content by either affecting their first or second-order beliefs. For example, users could be more afraid of legal repercussions of toxic posts (even though actual legal cases are extremely rare). The NetzDG could also change users' second-order beliefs about how acceptable other users find toxic content. Given the observational nature of our analysis, we will not be able to disentangle these hypotheses. Importantly, both of these mechanisms are in line with the effect of the NetzDG on online and ultimately offline behavior.

Offline Effects. To measure 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 high exposure to this type of content to be disproportionately affected by the NetzDG relative to places with a low exposure.
This intuition gives rise to the following empirical strategy:

$$y_{it} = \theta \cdot AfD\ Users\ p.c._i \times Post\ NetzDG_t + \mathbf{X_{it}'}\beta + \gamma_i + \delta_t + \epsilon_{it}, \tag{2}$$

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.[8] 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 starting in 2017q4 when the NetzDG became law. The regression includes a full set of municipality

[8]In Appendix Table A.9, we show that the results are robust to other variable transformations.

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). 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 if 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. Table A.2 plots the mean and standard deviation of a large number of municipality characteristics 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 crime 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.[9]

Synthetic Control Methods. Lastly, we investigate the effect of the NetzDG on the aggregate number of hate crime in Germany. To do so, we build a synthetic control 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 predictor the average of a dummy that indicates whether there was a change in measurement. As 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).

[9]In Appendix Table A.10, we show robustness for alternative levels of clustering.

5 Results

5.1 Did the NetzDG Reduce Online Toxicity?

We begin our analysis by providing evidence that the NetzDG reduced the toxicity of refugee-related social media content. Given that the main focus of this paper is on the impact of the NetzDG on hate crime, we report several of these findings in the online appendix.

Table 1 presents the results from estimating equation (1). Columns (1) and (2) show the result for users who posted highly toxic content before the NetzDG, while column (3) and (4) show the results for AfD users. As we show in Appendix Figure A.3, AfD Twitter followers are far more likely to post tweets with a toxicity above 0.8.[10] All specifications indicate a significant reduction in the toxicity of tweets after the NetzDG. The results also hardly change with the inclusion of user-specific linear time trends (see columns (2) and (4)).

The estimates for highly toxic users in column (1) suggests that the NetzDG was associated with a reduction in toxicity of around 32% relative to the mean. The magnitude for AfD users (see column (3)) is 4%. This should be unsurprising as the estimates in the first two columns focus on the most toxic users.[11]

In Appendix Table A.5 we provide a robustness check for different cutoffs of pre-period toxicity. Across all specifications, we find a reduction in online toxicity after the passing of the NetzDG. Appendix Table A.6 presents additional robustness exercises using the different measures of toxicity produced by Google's Perspective API. The effect is consistently significant and negative across almost all toxicity measures.

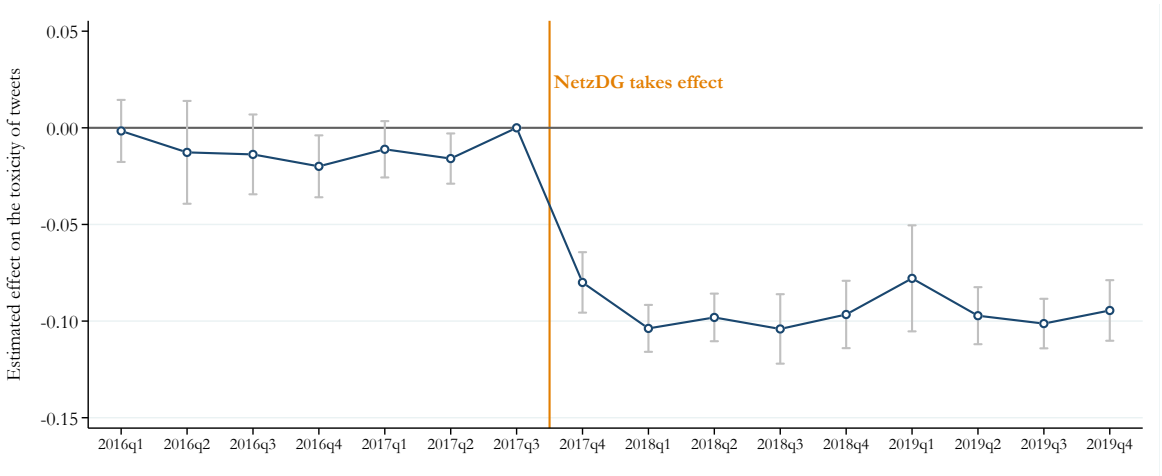
Figure 3 shows a dynamic event-study version of these specifications, which replaces the *Post* indicator variable with dummies for the quarters around when the NetzDG became active. Panel (a) shows the event study for highly toxic Twitter users, while Panel (b) shows the event study for AfD Twitter users. The figure suggests that the refugee-related tweets posted by AfD followers and other Twitter users had similar trends of toxicity up to 2017q3, which quickly and persistently turned negative with the start of the NetzDG becoming active in 2017q4.

[10] Many studies classify posts as hate speech if their toxicity is higher than 0.8 (ElSherief et al., 2018; Han and Tsvetkov, 2020; Vidgen et al., 2020).

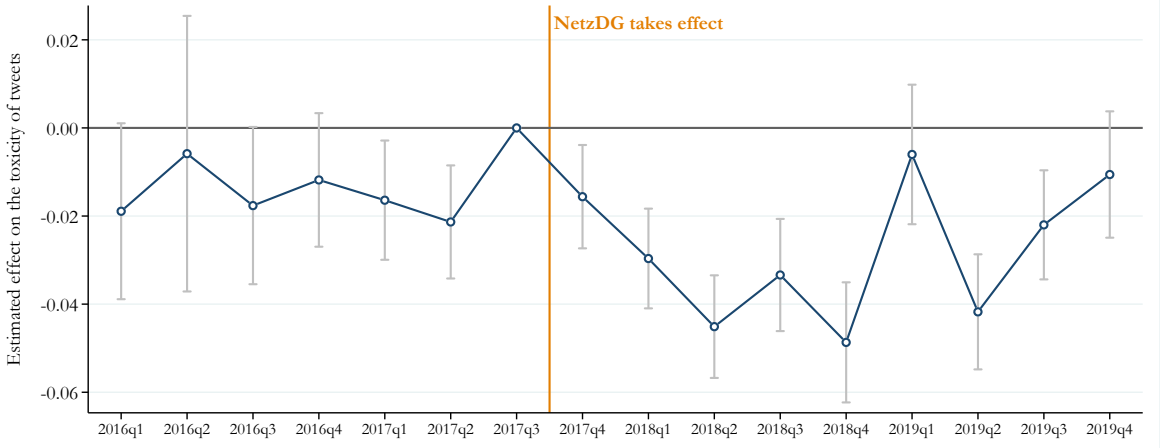
[11] Andres and Slivko (2021) find a reduction of around 2.5% relative to the mean (0.01 standard deviations) in the monthly volume of hateful tweets sent in Germany relative to Austria.

Figure 3: NetzDG and Online Toxicity of Refugee-related Content

(a) Highly Toxic Users



(b) AfD Followers



Notes: Panels A and B plot the coefficients from event study versions of Equation (1). In Panel A, we define $Toxic\ User_u$ equal to 1 if a user was in the top decile of toxicity pre-NetzDG, and 0 otherwise. In Panel B, we define $Toxic\ User_u$ equal to 1 if a user 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.

Table 1: Regression Estimates: NetzDG and Refugee-related Online Toxicity

	Dep. var.: Toxicity Measures			
	(1)	(2)	(3)	(4)
Highly Toxic Users × Post	-0.085***	-0.075***		
	(0.004)	(0.006)		
AfD followers × Post			-0.016***	-0.018***
			(0.003)	(0.004)
User FE	Yes	Yes	Yes	Yes
Day FE	Yes	Yes	Yes	Yes
Linear Time Trend		Yes		Yes
Observations	275,054	275,054	275,054	275,054
Pre-Period Mean of DV	0.39	0.39	0.39	0.39
R^2	0.28	0.34	0.28	0.34

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) $Toxic\ User_u$ is an indicator variable equal to 1 if a users' tweets before the NetzDG were on average above the 75th percentile. In columns (3) and (4), $Toxic\ User_u$ is an indicator variable that is equal to 1 if a Twitter user follows the AfD's account. All regressions control for user and day fixed effects. Columns (2) and (4) additionally control for user-specific linear time trends. Robust standard errors in parentheses are clustered by user. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

In Appendix Figure A.4 we provide additional evidence for the effectiveness of the NetzDG by investigating the overall toxicity of Twitter content. Panel (a) plots the average toxicity of Tweets sent by highly toxic users. Similar to the results for refugee-related content, we again observe a significant reduction in online toxicity after the NetzDG. In panel (b), we additionally show that the NetzDG did not have any effect on the total number of tweets send by highly toxic users.

It is again worth noting that we cannot disentangle whether these findings are driven by platforms deleting an increasing number of hateful tweets after the implementation of the NetzDG or due to a deterrence effect leading users to self-censor. However, taken together, they do suggest that the NetzDG was associated with a reduction of the toxicity of German far-right refugee-related social media content, which is what matters for our analysis. In the next section, we study whether the NetzDG-induced drop in

hateful online rhetoric also affected real-life violence.

5.2 Online Content Moderation and Hate Crimes

Baseline estimates. Table 2 shows our main results for the effect of the NetzDG on anti-refugee hate crime. Column (1) contains estimates of our baseline specification using Equation (2), controlling only for the log number of inhabitants interacted with the *Post* indicator to control for any changes in hate crimes due to population differences. In the following columns, we add controls for some of the most relevant potential confounders. Column (2) adds Facebook users per capita in a municipality to account for any changes in anti-refugee incidents that could be explained by unobservable counfounders that correlate with a municipality's affinity ot social media. In a similar spirit, we add a control for the access to broadband internet in column (3). In column (4), we additionally control for the vote share of the AfD at the municipality level. This control will account for changes in anti-refugee incidents around the time of the NetzDG that can be explained by the overall support for the AfD. Finally, in column (5) we include a wealth of additional control variables (see Appendix A.2. for details) all of which we interact with the *Post* indicator.

Including additional control variables has little impact on the magnitude, sign, and statistical significance of our main estimate. Importantly, the coefficients capturing a town's general degree of Facebook or internet penetration are not consistently statistically significant and quantitatively small. In other words, after accounting for the exposure to far-right Facebook usage, a town's social media penetration has little impact on its responsiveness to the NetzDG. Finally, controlling for the AfD vote share—which captures ways in which far-right support might affect a municipality's response to the NetzDG—leaves our main coefficient of interest virtually unchanged.

In our preferred specification—column (4), which controls for population, and voting and media consumption behavior, the -0.009 point estimate indicates that a one standard deviation increase in AfD Facebook users per capita results in a -0.9 percentage point (relative) reduction in quarterly 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. This estimate also seem plausible given the 4% reduction in hateful online content we identified in the previous section.

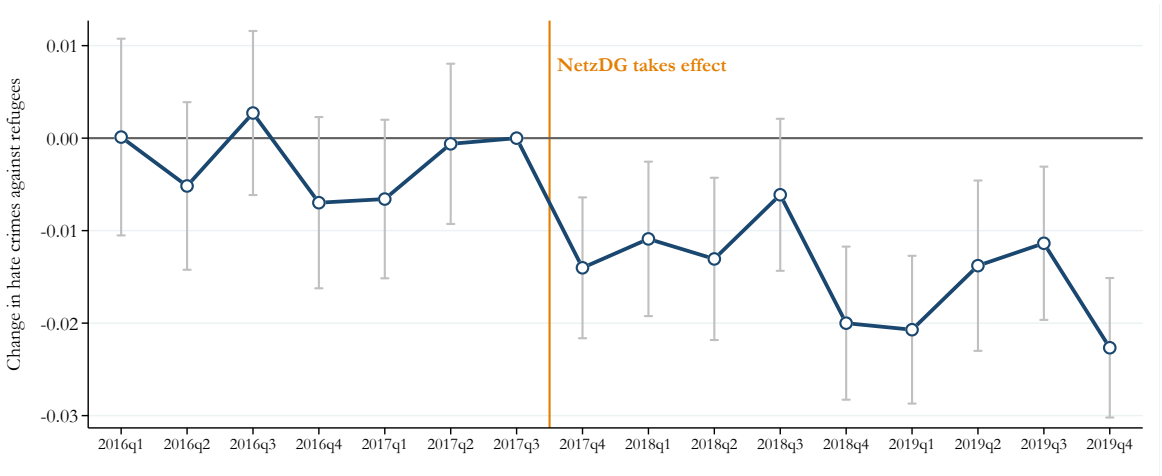
Table 2: Regression Estimates: Effect of NetzDG on Hate Crime

	Dep. var.: Asinh(Anti-Refugee Hate Crimes)				
	(1)	(2)	(3)	(4)	(5)
AfD Facebook users p.c. (std) × Post	-0.012***	-0.012***	-0.012***	-0.009***	-0.008***
	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)
Facebook users p.c (std) × Post		0.003*	0.003*	0.002	0.003
		(0.002)	(0.002)	(0.002)	(0.002)
Broadband internet (std) × Post			0.010***	0.005	0.001
			(0.003)	(0.003)	(0.004)
AfD vote share (std) × Post				-0.012***	0.031***
				(0.004)	(0.012)
Ln(Pop.) × Post	Yes	Yes	Yes	Yes	Yes
Municipality FE	Yes	Yes	Yes	Yes	Yes
Year-Quarter FE	Yes	Yes	Yes	Yes	Yes
All Controls (19) × Post					Yes
Observations	71,456	71,456	71,456	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 (2), 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. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Event study. As with any difference-in-differences design, we require that municipalities with different prior exposure to right-wing content would have followed similar trends in the absence of the increased content moderation prompted by the NetzDG. While this assumption is inherently untestable, we can provide some evidence in support of it. In particular, we estimate a quarterly event study to show that municipalities with many and few AfD users followed a very similar trajectory before the NetzDG. The event study also allows us to analyze the dynamics of the treatment effects. Figure 4 visualizes the coefficients from the event study regression relative to the 3rd quarter of 2017 (the quarter before the NetzDG became law). We find no evidence for pre-existing trends in this specification, i.e., all 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. This negative effect appears to be persistent and stable over the two years following the NetzDG.

Figure 4: Event Study Hate Crime



Notes: This figure plots the coefficients from running an event study version of regression Equation (2). 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.

Heterogeneous effects. If the effect of an increase in content moderation depends on the exposure to hateful content, we would expect to see heterogeneity of our estimates by the usage intensity of the AfD Facebook page. In other words, even if two municipalities have the same number of AfD Facebook users per capita, we expect to see a bigger impact of the NetzDG in the municipality in which the AfD users are more active. In Table 3, we explore this possibility by including different measures of usage intensity in the regressions. In particular, we measure the usage intensity of the AfD's Facebook page using the average number of posts, comments, likes, and shares sent by each AfD user in a given municipality before the passing of the NetzDG. Note that these regressions are only estimated for municipalities for which we can identify at least one AfD user.

The results in Table 3 suggest that the effect of the NetzDG is stronger in municipalities in which users were more actively interacting with the AfD's Facebook page. This holds independent of the measure of usage intensity we are using. 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. We also find that the effect of the number of AfD Facebook users per capita in this subset of municipalities with at least one AfD

user is even stronger.

These findings also lend further support to the underlying assumption of our empirical strategy because they show that both the extensive and intensive margin of AfD Facebook usage matters. Any alternative explanation would have to account for the fact that we see a larger reduction in hate crimes in municipalities that have similar numbers of AfD users but more active users, which makes it less likely that we are capturing unobservable confounding variables.

Table 3: 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)
Ln(Pop.) × Post	Yes	Yes	Yes	Yes
Municipality FE	Yes	Yes	Yes	Yes
Quarter FE	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 (2) 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. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Alternative Explanations. As with any difference in difference estimate, we need to assume that there is no other shock at the same time that differentially affects areas

with many AfD Users. Two possible candidates for such shocks could be 1) the end of the refugee crisis, and 2) the 2017 federal election. We discuss these events in turn.

First, the finding cannot easily be explained by some form of mean reversion in the number of anti-refugee incidents due to the end of the refugee crisis in Germany. 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. As such the effect, we find occurs over a year after this important demarcation point of the refugee crisis. It is further worth noting, that our exposure measure is largely uncorrelated with the number of refugees in each municipality (see Appendix Table A.2. As a result, the inclusion of this control makes no difference for our estimates (see column 5 Table 2).

Second, also the 2017 federal election seems highly unlikely to drive our findings as we include controls for the electoral results of all major German parties in our regressions. This again makes hardly any difference for the magnitude and significance of our findings. Further, the fact that the coefficient for the AfD vote share is positive in the specification with all controls contradicts the idea that the end of the election period was associated with a drop in the number of anti-refugee incidents.

Lastly, also any other event that occurred at the time of the passage of the NetzDG is unlikely to bias our estimates, as would be surprising if any such event affects municipalities with many AfD Facebook users but at the same doesn't affect municipalities with many Facebook users or AfD voters.

Robustness. To probe the robustness of our findings, we perform six additional robustness checks. First, in Online Appendix Table A.8 we show that with the exception of arson there is an effect of the NetzDG on all categories of anti-refugee hate crimes we are considering (i.e., assault, demonstration, suspected attacks, and other (miscellaneous) property attacks). We observe the strongest response to the NetzDG for assaults and other property attacks. This makes it unlikely we are capturing changes in reporting of minor incidents.

Next, Table 4 presents a battery of additional robustness exercises. In column (2), we show that our findings are robust 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 be 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). In column (3), we exclude January and February 2016 from our data, which contain the largest spike in anti-refugee incidents in our data. This leaves our results completely unchanged. Similarly, our 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, our results remain highly statistically significant.

Table 4: Robustness

	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***	-0.009***	-0.010***	-0.017***	-0.009***	-0.018***
	(0.002)	(0.002)	(0.002)	(0.005)	(0.003)	(0.003)
Ln(Pop/) × Post	Yes	Yes	Yes	Yes	Yes	Yes
AfD vote share × 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
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes
Quarter FE	Yes	Yes	Yes	Yes	Yes	Yes
Fed. State × Quarter FE		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 (2) 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. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Third, Table A.9 shows that our estimates are robust independently of the functional form of the dependent and independent variables we are using. In particular, we explore transformations of the dependent variable (refugee attacks) in inverse hyperbolic sine (baseline), counts, or the log number of refugee incidents per capita. Neither of these changes makes any difference to our findings (see column (1-3)). In column (4-6), we 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 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 rules out that our findings could be driven by heterogeneous treatment effects in

our two-way fixed effects estimation (De Chaisemartin and D'Haultfoeuille, 2022), as our results also hold in this dummy specification.

Fourth, we repeat our analysis based on the number of AfD Twitter followers in a municipality. In Appendix Table A.11 we show that the overall results are virtually identical if we use this alternative measure of exposure to the NetzDG. Appendix Figure A.6 also presents the corresponding event study estimates.

Fifth, we perform a leave-one-out analysis in which we exclude one municipality at a time. The results are shown in Appendix Figure A.5. The estimates are highly stable throughout as such our findings do not appear to be driven by outliers.

Finally, Table A.10 shows that our estimates remain statistically significant irrespective of the level of clustering of the standard errors. More specifically, we show that our main 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.

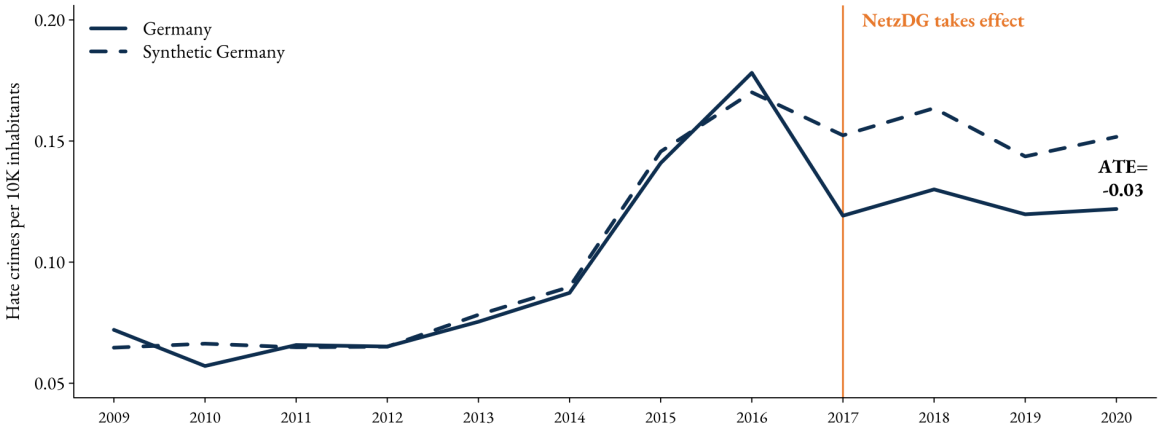
5.3 Synthetic Control Estimates

The previous findings suggest that the NetzDG resulted in a relative decrease in offline violence against refugees in Germany between municipalities with differential exposure to anti-refugee rhetoric online. To further probe these results, and provide a sense of the aggregate effect of the policy beyond anti-refugee incidents, we estimate a synthetic control for Germany as described in Section 4.

We report the main estimates from this exercise in Figure 5. This figure shows that the number of hate crimes per 10,000 inhabitants in the synthetic Germany built from the 21 donor countries closely track 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 crime 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.

In Appendix Appendix A.3.3, we provide several robustness checks for these synthetic control estimates. The results are robust to changing the interpolation assumptions, outcome variables, time periods, and donor countries. Moreover, we can reject the null hypothesis that the ATE is non-negative with a p-value of 0.045, constructed based on in-space placebo tests as in Abadie et al. (2010). Appendix A.3.3 also presents additional information, such as the weights used to construct the synthetic

Figure 5: 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 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 pre-period.

control and the pre-period balance of the predictors. Hence, the results suggest that the NetzDG contributed to reduce the aggregate number of hate crimes in Germany relative.

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 attempts of these platforms to moderate this content. By leveraging this unique quasi-experiment, this study 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. As such, we believe our findings should

best be interpreted as a useful starting point for understanding the online and offline effects of online content moderation.

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.

Acemoglu, D., A. Ozdaglar, and J. Siderius (2021). Misinformation: Strategic Sharing, Homophily, and Endogenous Echo Chambers. Technical report, National Bureau of Economic Research.

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.

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.

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).

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

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.

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 of the United States of America*, 201706588.

Braghieri, L., R. Levy, and A. Makarin (2022). Social Media and Mental Health.

Bundesamt für Justiz (2019). Federal Office of Justice Issues Fine against Facebook. `https://www.bundesjustizamt.de/DE/Presse/Archiv/2019/20190702_EN.html`. Accessed: 2021-09-30.

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.

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.

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 (2022). Two-Way Fixed Effects and Differences-In-Differences With Heterogeneous Treatment Effects: A Survey. Technical report, National Bureau of Economic Research.

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.

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.

ElSherief, M., V. Kulkarni, D. Nguyen, W. Y. Wang, and E. Belding (2018). Hate Lingo: A Target-Based Linguistic Analysis of Hate Speech in Social Media. In *Proceedings of the International AAAI Conference on Web and Social Media*, Volume 12.

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

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. Müller, and C. Schwarz (2021). The Effect of Social Media on Elections: Evidence From the United States. *NBER Working Papper*.

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

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

Han, X. and Y. Tsvetkov (2020). Fortifying Toxic Speech Detectors Against Veiled Toxicity. *arXiv preprint arXiv:2010.03154*.

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

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.

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.

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*.

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 (2022a). The effects of online content moderation: Evidence from president trump's account deletion. *Available at SSRN 4296306*.

Müller, K. and C. Schwarz (2022b). From Hashtag to Hate Crime: Twitter and Anti-Minority Sentiment. *Forthcoming American Economic Journal: Applied Economics*.

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

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.

Vidgen, B., S. Hale, S. Staton, T. Melham, H. Margetts, O. Kammar, and M. Szymczak (2020). Recalibrating Classifiers for Interpretable Abusive Content Detection. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pp. 132–138.

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.

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

A Online Appendix

A.1. 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.[12] 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:

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

We interpret the implementation of the NetzDG as a marginal increase in the expected regulatory penalty; $d\tau > 0$.[13] 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 (A.1) hold, the amount of surviving hateful content on the platform decreases in response to an increase in fines; $dT^H/d\tau < 0$.[14]

[12] 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.

[13] While the NetzDG was the clearest shift in regulatory incentives for content moderation, in practice fines have been small. For example, in 2019, Germany fined Facebook €2 million for violating the NetzDG law (Bundesamt für Justiz, 2019).

[14] To see why, rewrite problem (A.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 $\widetilde{\pi}'' < 0$.

A.2. Additional Details on the Data

Table A.1: 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 User 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 Greens	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.2: 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 User 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 Greens	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

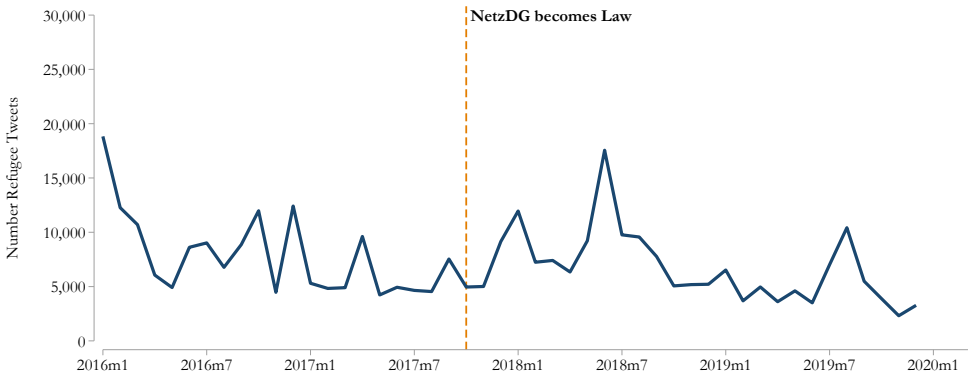
Notes: This table displays the mean, standard deviation, of the variables used in the municipality-year-quarter panel, split by quartiles of AfD Facebook users per capita (the "exposure" variable in the difference-in-differences analysis).

Table A.3: Summary Statistics Toxicity Refugee Tweets

Variable	Mean	SD	p50	Min	Max	N
Toxicity Measures						
Toxicity	0.41	0.22	0.00	0.41	1.00	346,167
Sev. Toxicity	0.31	0.24	0.00	0.29	1.00	346,167
Identity Attack	0.52	0.25	0.00	0.52	1.00	346,167
Insult	0.35	0.20	0.00	0.35	1.00	346,167
Profanity	0.22	0.21	0.00	0.12	1.00	346,167
Threat	0.41	0.29	0.00	0.25	1.00	346,167
User Measures						
AfD Twitter Followers	0.28	0.45	0.00	0.00	1.00	346,167
Party Twitter Followers	0.49	0.50	0.00	0.00	1.00	346,167
Pre-Period Tox $\geq$ 50pct	0.53	0.50	0.00	1.00	1.00	295,695
Pre-Period Tox $\geq$ 75pct	0.25	0.43	0.00	0.00	1.00	295,695
Pre-Period Tox $\geq$ 90pct	0.10	0.30	0.00	0.00	1.00	295,695
Pre-Period Tox $\geq$ 95pct	0.05	0.22	0.00	0.00	1.00	295,695

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

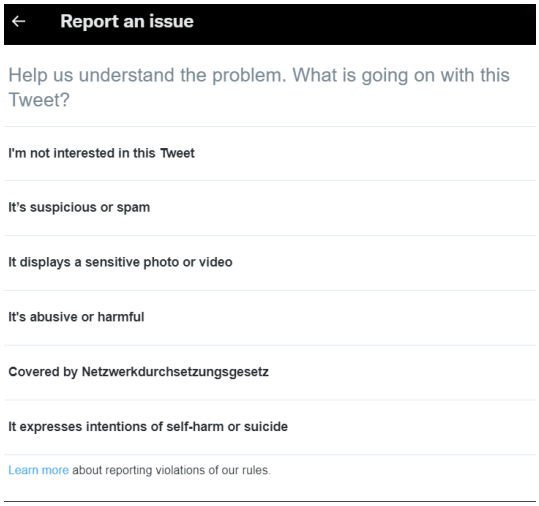
Figure A.1: Time Series Refugee Tweets



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

Figure A.2: How Users Twitter 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.

Table A.4: OSCE Members and Data Filters

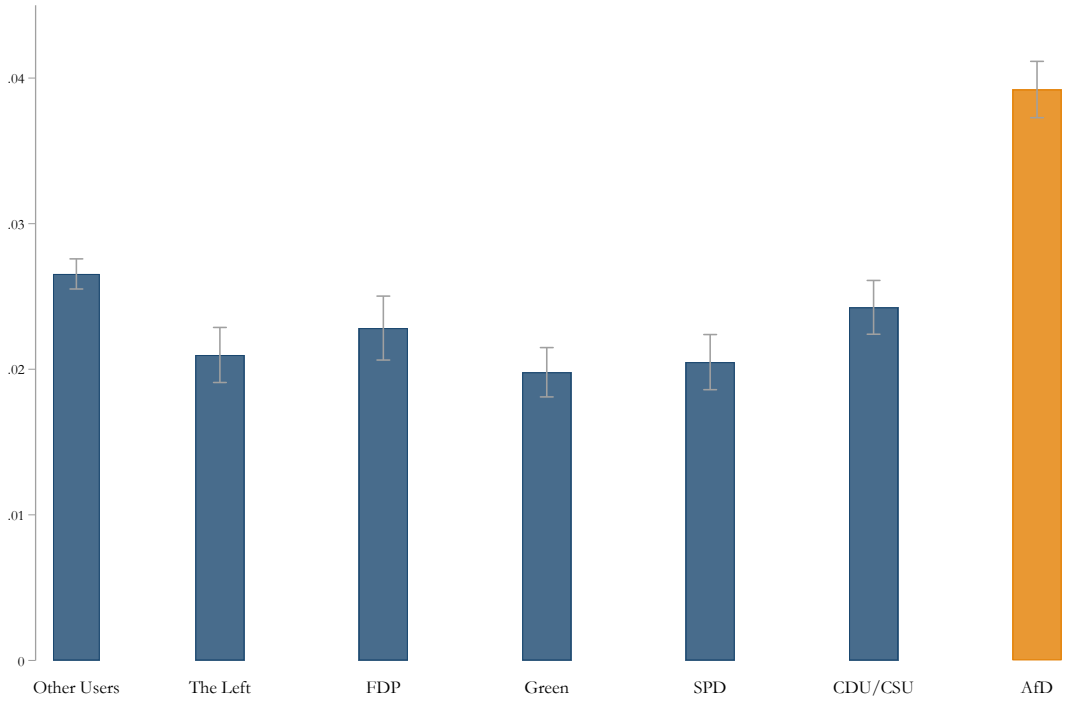
OSCE State	No data 2009-2020	Microstate	Data changes 2017-2020	7+ missings 2009-2020	End gaps
Albania				×	×
Andorra		×			
Armenia				×	×
Austria					
Azerbaijan				×	×
Belarus				×	×
Belgium					
Bosnia and Herzegovina					
Bulgaria					
Canada			×		
Croatia					
Cyprus					
Czech Republic					
Denmark					
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.

A.3. Additional Results

A.3.1 Additional Results for the Toxicity of Tweets

Figure A.3: Toxicity by Party in Pre-Period



Notes: The figure shows bar graphs with the frequency of tweets with a toxicity larger than 0.8 depending on which German party users follow before the passing of the NetzDG.

Table A.5: Robustness: Threshold of Pre-Period Toxicity

	Dep. var.: Toxicity Measures			
	(1)	(2)	(3)	(4)
Pre-Period Tox $\geq$ 50pct $\times$ Post	-0.080***			
	(0.002)			
Pre-Period Tox $\geq$ 75pct $\times$ Post		-0.085***		
		(0.004)		
Pre-Period Tox $\geq$ 90pct $\times$ Post			-0.131***	
			(0.005)	
Pre-Period Tox $\geq$ 95pct $\times$ Post				-0.179***
				(0.006)
User FE	Yes	Yes	Yes	Yes
Day FE	Yes	Yes	Yes	Yes
Observations	275,054	275,054	275,054	275,054
Pre-Period Mean of DV	0.39	0.39	0.39	0.39
R^2	0.28	0.28	0.28	0.28

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). $Toxic\ User_u$ 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. All regressions control for user and day fixed effects. Robust standard errors in parentheses are clustered by user. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

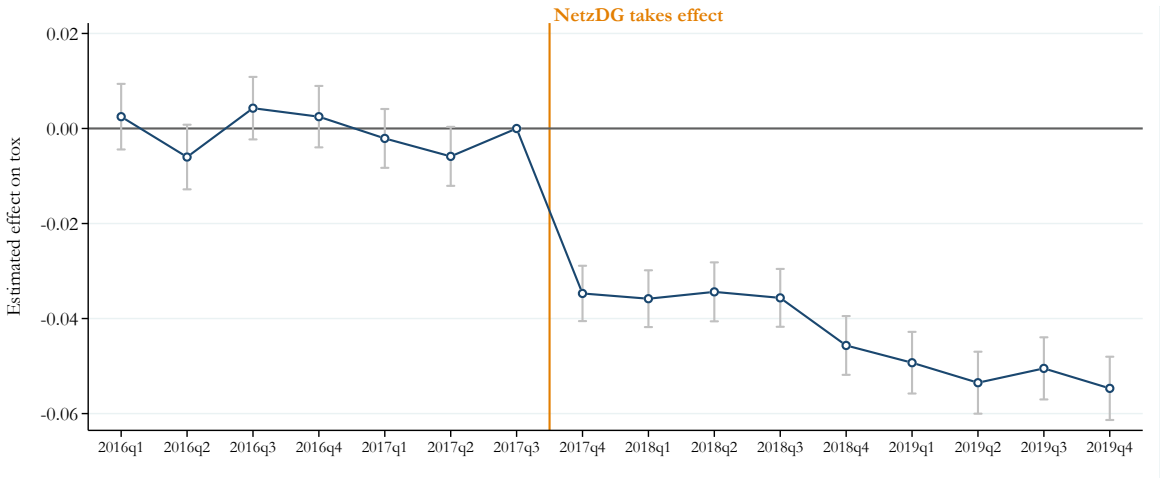
Table A.6: Robustness: Toxicity Measures

	Dep. var.: Toxicity measured by:					
	Toxicity	Severe Toxicity	Identity Attack	Insult	Profanity	Threat
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Highly Toxic Users						
Highly Toxic Users × Post	-0.085***	-0.086***	-0.084***	-0.074***	-0.073***	-0.061***
	(0.004)	(0.004)	(0.004)	(0.003)	(0.003)	(0.006)
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Day FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	275,054	275,054	275,054	275,054	275,054	275,054
Pre-Period Mean of DV	0.39	0.30	0.51	0.33	0.21	0.41
R^2	0.28	0.27	0.26	0.27	0.24	0.29
Panel B: AfD Followers						
AfD followers × Post	-0.016***	-0.017***	-0.023***	-0.017***	-0.019***	0.003
	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)	(0.004)
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Day FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	275,054	275,054	275,054	275,054	275,054	275,054
Pre-Period Mean of DV	0.39	0.30	0.51	0.33	0.21	0.41
R^2	0.28	0.27	0.26	0.27	0.24	0.29

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 Twitter 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. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Figure A.4: NetzDG and Overall Online Toxicity

(a) Average Toxicity All Tweets



Notes: The figure plot the coefficients from event study versions of Equation (1). The dependent variable is the toxicity of all tweets send 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 A.7: Regression Estimates: NetzDG and Overall Online Toxicity

	Dep. var.:			
	Toxicity Measure		*Asinh(Nr. Tweets)*	
	(1)	(2)	(3)	(4)
Highly Toxic User × Post NetzDG	-0.024***	-0.014***	0.098***	0.005
	(0.001)	(0.001)	(0.015)	(0.016)
User FE	Yes	Yes	Yes	Yes
Quarter FE	Yes	Yes	Yes	Yes
Linear Time Trend		Yes		Yes
Observations	610,059	610,059	627,610	627,610
Pre-Period Mean of DV	0.17	0.17	0.17	0.17
R^2	0.55	0.63	0.54	0.74

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), the dependent variable is the average toxicity of tweets send by user i in quarter t. In columns (3) and (4), the dependent variable is the inverse hyperbolic sine of the number of tweets send by user i in quarter t. $Toxic\,User_u$ is an indicator variable equal to 1 if a users' tweets before the NetzDG were on average above the 75th percentile. 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. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

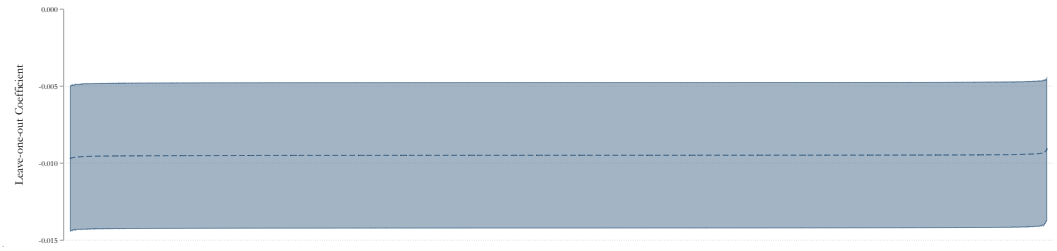
A.3.2 Additional Results for Hate Crimes

Table A.8: 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)
Ln(Pop/) × Post	Yes	Yes	Yes	Yes	Yes	Yes
AfD vote share × 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
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes
Quarter FE	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 (2) 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 A.5: 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 A.9: 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***	-0.025***	-0.007***			
	(0.002)	(0.005)	(0.002)			
High AfD Usage × Post				-0.026***	-0.081***	-0.020***
				(0.007)	(0.023)	(0.005)
Ln(Pop/) × Post	Yes	Yes	Yes	Yes	Yes	Yes
AfD vote share × 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
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes
Quarter FE	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 (2) 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 A.10: 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)
Ln(Pop/) × Post	Yes	Yes	Yes	Yes
AfD vote share × Post	Yes	Yes	Yes	Yes
Facebook users p.c × Post	Yes	Yes	Yes	Yes
Broadband internet × Post	Yes	Yes	Yes	Yes
Municipality FE	Yes	Yes	Yes	Yes
Quarter FE	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

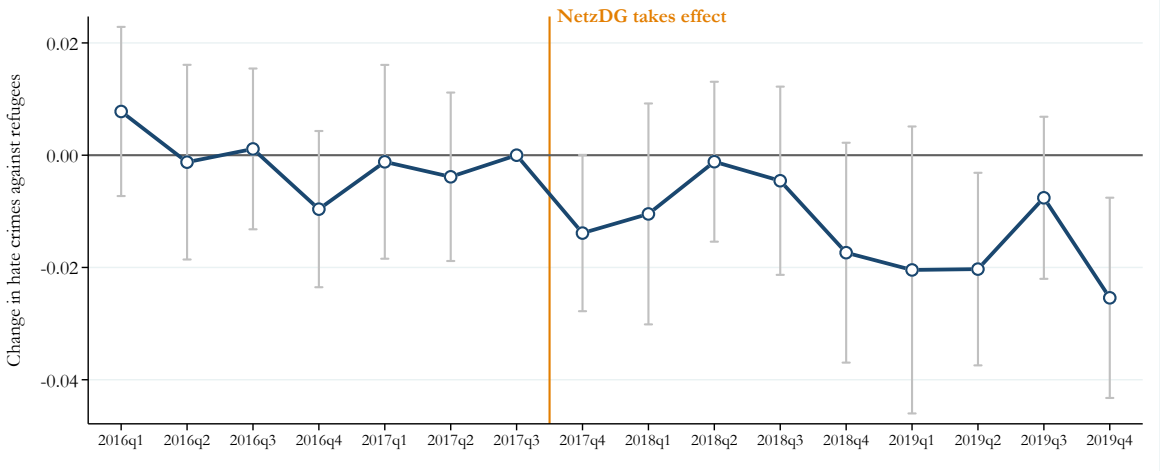
Notes: This table presents the results of estimating municipality-quarter-level regressions as in Equation (2) where the dependent variable is the inverse hyperbolic sine of the number of anti-refugee hate crimes. *AfD Facebook users p.c. (std)* is the number of AfD Facebook followers per capita, standardized 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. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.11: Robustness: Social Media Exposure measured with Twitter

	Dep. var.: Asinh(Anti-Refugee Hate Crimes)				
	(1)	(2)	(3)	(4)	(5)
AfD Twitter Follower p.c. (std) × Post	-0.012**	-0.013**	-0.012**	-0.011**	-0.011**
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Facebook users p.c (std) × Post		0.003*	0.003*	0.002	0.003
		(0.002)	(0.002)	(0.002)	(0.002)
Broadband internet (std) × Post			0.010***	0.004	0.000
			(0.004)	(0.003)	(0.004)
AfD vote share (std) × Post				-0.014***	0.029**
				(0.004)	(0.012)
Ln(Pop.) × Post	Yes	Yes	Yes	Yes	Yes
Municipality FE	Yes	Yes	Yes	Yes	Yes
Year-Quarter FE	Yes	Yes	Yes	Yes	Yes
All Controls (19) × Post					Yes
Observations	71,456	71,456	71,456	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 (2), 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. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

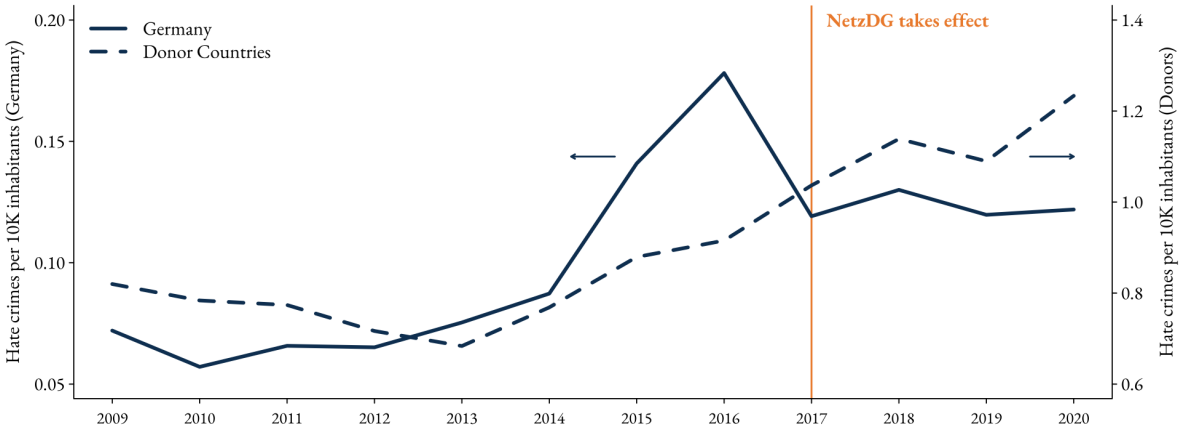
Figure A.6: Event Study Hate Crime (Twitter Exposure)



Notes: This figure plots the coefficients from running an event study version of regression Equation (2). 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.

A.3.3 Synthetic Control

Figure A.7: Evolution of Hate Crimes in Germany vs. Donor Countries



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

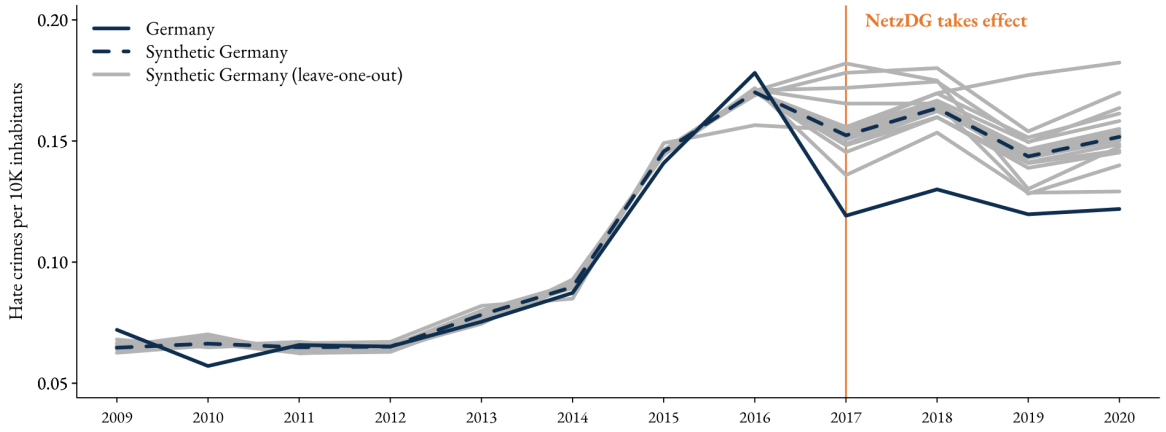
Table A.12: Hate Crimes Predictor Means

Variable	Germany		Donors	OECD	OSCE
	Real	Synthetic			
Hate crimes per 10K inhabitants 2009	0.07	0.04	0.74	0.97	0.67
Hate crimes per 10K inhabitants 2010	0.06	0.04	0.68	0.86	0.6
Hate crimes per 10K inhabitants 2011	0.07	0.04	0.69	0.85	0.58
Hate crimes per 10K inhabitants 2012	0.07	0.04	0.6	0.78	0.55
Hate crimes per 10K inhabitants 2013	0.08	0.06	0.62	0.71	0.51
Hate crimes per 10K inhabitants 2014	0.09	0.07	0.67	0.82	0.55
Hate crimes per 10K inhabitants 2015	0.14	0.12	0.79	0.96	0.65
Hate crimes per 10K inhabitants 2016	0.18	0.14	0.86	1.03	0.68
Measure change 2009-2016	0	0.11	0.11	0.07	0.05

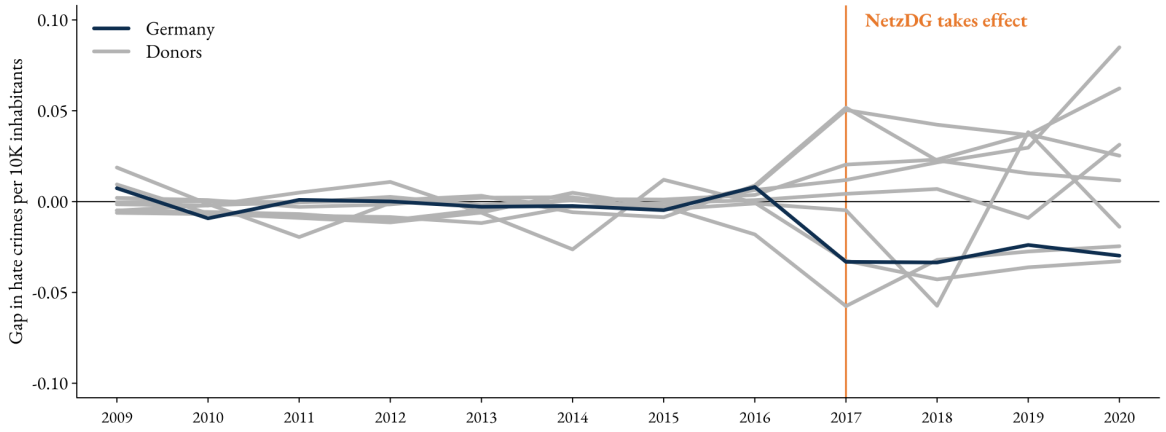
Notes: This table presents the means of the predictor variables for Germany and the synthetic Germany, as well as the simple mean among the donor, OECD, and OSCE countries.

Figure A.8: Leave-One-Out and In-Space Placebos

(a) Germany vs. Synthetic Control

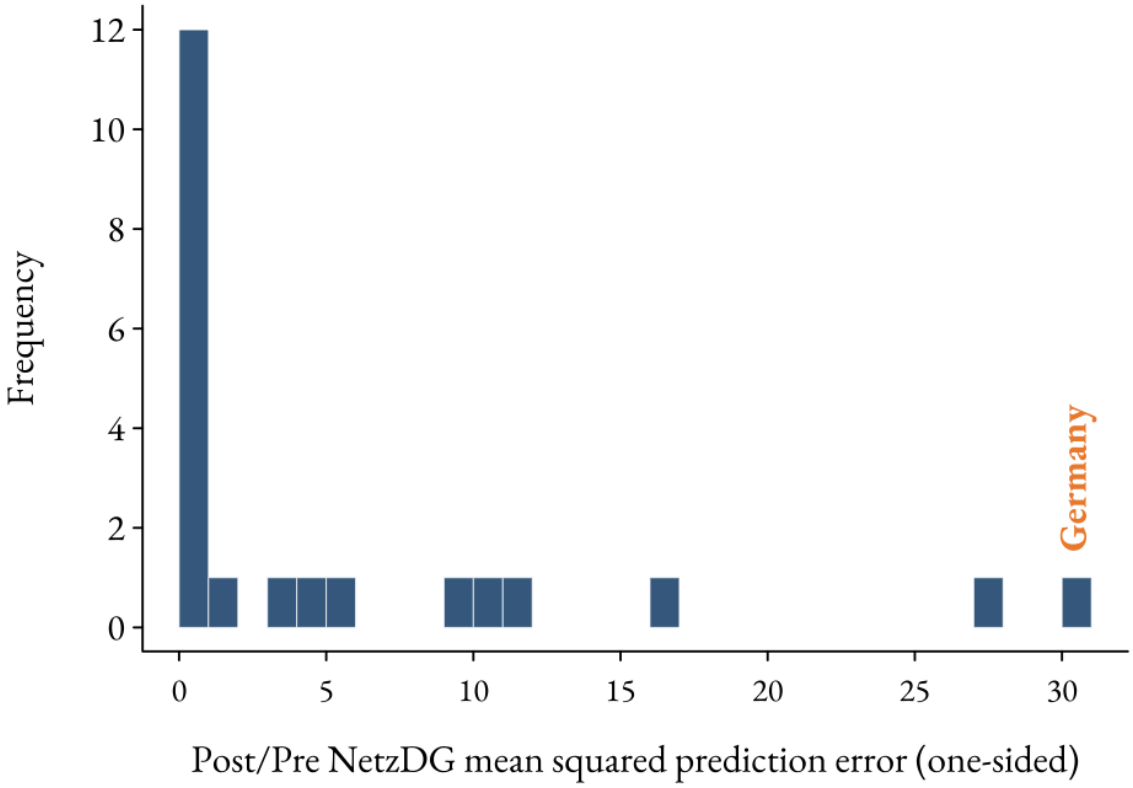


(b) Gaps Between Observed and Synthetic Hate Crimes



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 A.9: 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 A.13: 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

Table A.14: Robustness to Alternative Specifications

Specification	ATE	p-value (one-sided)	p-value (two-sided)	Donors	Pre-NetzDG RMPSE
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
Asinh	-0.012	0.227	0.682	21	0.006
Levels	-0.047	0.136	0.318	21	0.012
Hate crimes per refugee	-0.335	0.455	0.636	21	0.064
Violent and non violent hate crimes	-0.198	0.364	0.773	21	0.115
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 ATE)	-0.014	0.19	0.524	20	0.006
Leave-one-out (min ATE)	-0.048	0.048	0.238	20	0.009
OECD	-0.067	0.067	0.133	14	0.007
OSCE	-0.056	0.161	0.226	30	0.004

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 ATE 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.