DISCUSSION PAPER SERIES

DP17554

THE EFFECT OF CONTENT MODERATION ON ONLINE AND OFFLINE HATE: EVIDENCE FROM GERMANY'S NETZDG

Carlo Schwarz, Rafael Jiménez Durán and

POLITICAL ECONOMY

ISSN 0265-8003

THE EFFECT OF CONTENT MODERATION ON ONLINE AND OFFLINE HATE: EVIDENCE FROM GERMANY'S NETZDG

Carlo Schwarz, Rafael Jiménez Durán and

Discussion Paper DP17554
Published 01 October 2022
Submitted 01 October 2022

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

This Discussion Paper is issued under the auspices of the Centre’s research programmes:

- Political Economy

Any opinions expressed here are those of the author(s) and not those of the Centre for Economic Policy Research. Research disseminated by CEPR may include views on policy, but the Centre itself takes no institutional policy positions.

The Centre for Economic Policy Research was established in 1983 as an educational charity, to promote independent analysis and public discussion of open economies and the relations among them. It is pluralist and non-partisan, bringing economic research to bear on the analysis of medium- and long-run policy questions.

These Discussion Papers often represent preliminary or incomplete work, circulated to encourage discussion and comment. Citation and use of such a paper should take account of its provisional character.

THE EFFECT OF CONTENT MODERATION ON ONLINE AND OFFLINE HATE: EVIDENCE FROM GERMANY'S NETZDG

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 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 statistically significant reduction in toxic posts by far-right social media users. Further, we show that the NetzDG reduced anti-refugee hate crimes in towns with more far-right Facebook users. Together, these findings suggest that online content moderation can curb online hate speech and offline violence.

JEL Classification: L82, J15, O38

Keywords: Refugees, Germany

Carlo Schwarz - carlo.schwarz@unibocconi.it
Bocconi University and CEPR

Rafael Jiménez Durán - rjimenez@ssrc.org
Social Science Research Council

 - kmueller@nus.edu.sg
National University of Singapore Business School

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

Rafael Jiménez Durán[*] Karsten Müller[†] Carlo Schwarz[‡]

September 26, 2022

Abstract

Social media companies are under scrutiny for the prevalence of hateful content on their platforms, but there is scarce empirical evidence on the consequences of regulating such content. We study this question 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 statistically significant reduction in toxic posts by far-right social media users. Further, we show that the NetzDG reduced anti-refugee hate crimes in towns with more far-right Facebook users. Together, these findings suggest that online content moderation can curb online hate speech and offline violence.

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

[*] Social Science Research Council, Stigler Center, rjimenez@ssrc.org

[†] 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, 2019; 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 the effectiveness of content moderation policies by focusing on the first legal change aimed at increasing the moderation efforts of social media platforms: the German "Netzwerkdurchsetzungsgesetz" (Network Enforcement Act, henceforth NetzDG). This policy was enacted on September 1, 2017 in response to a spike in online hate speech that coincided with the influx of more than one million refugees into Germany during the 2015-2016 refugee crisis. The NetzDG marks a unique and unprecedented legal change in Germany that introduced large penalties for social media platforms—up to 50 million euros—for failing to promptly remove hateful content.[1] As such, the law drastically changed social media providers' incentives to remove hateful content.

In this paper, we investigate whether increased content moderation efforts induced by the NetzDG indeed decreased online and offline hatred targeting minorities. In particular, we use a difference-in-difference methodology exploiting differential exposure to hateful social media content to study the effects of the NetzDG's enactment. Following Müller and Schwarz (2021), we proxy for exposure to anti-refugee content online using

[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.

the Twitter and Facebook accounts of the Alternative for Germany ("Alternative für Deutschland", henceforth AfD). At the time the NetzDG became effective, the AfD 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) far more Facebook followers than any other German party.

As a first stage, we provide evidence that the NetzDG was indeed followed by a decrease in toxic content on social media, as measured by Google's toxicity score—a measure commonly used in industry applications and as benchmark in academic studies. We compare the toxicity score of a large sample of Tweets posted by Twitter followers of the AfD relative to similar Tweets posted by followers of other German parties.[2] Intuitively, the higher toxicity of AfD users' posts leaves them more exposed to online content moderation aimed at hate speech. In line with an effect of the NetzDG, we observe a significant reduction in the toxicity of Tweets of AfD followers after the implementation of the law. We also show that this finding is consistent with a simple theoretical framework that models content moderation as a quality decision for platforms and the NetzDG as a tax on unmoderated content.

For our main analysis, we then 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 find empirically, one would expect a larger 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 a one standard deviation higher number of AfD followers per capita saw a -0.8 percentage point reduction in the number of anti-refugee incidents.

In addition, we also investigate the intensive margin of far-right Facebook usage. We find that the intensity with which users interact with the AfD's Facebook page (as measured by posts, likes, comments, or shares) is associated with a stronger reduction of anti-refugee hate crimes, over and above what is predicted by the number of AfD followers. For example, municipalities with a 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.

[2] This analysis is conducted on Twitter because Facebook, unfortunately, does not allow us to collect the posts of private users.

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 evidence that municipalities with different levels of AfD followers had identical trends in the period leading up to the enactment of the NetzDG, consistent with parallel trends. 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, nor by strong support for the AfD in the 2016 federal election. If anything, the coefficients for these variables appear to be positive, although they are mostly statistically insignificant. Furthermore, the main results remain unchanged if we consider alternative variable transformations, standard errors, more restrictive fixed effects, and sub-samples for our analysis. Lastly, we conduct placebo checks by investigating the impact of the NetzDG on crimes that are less likely to be driven by online hate speech (e.g., theft or drugs). In line with our expectation, we find no significant reduction in these crimes, which also alleviates concerns that our findings could be driven by an overall reduction in the crime rate in municipalities with many AfD followers.

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, 2019; 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 can reduce real-life violence.

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 implies that the most sizeable welfare effects of content moderation could be due to its impact on out-of-platform outcomes, such as hate crimes. Our 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.

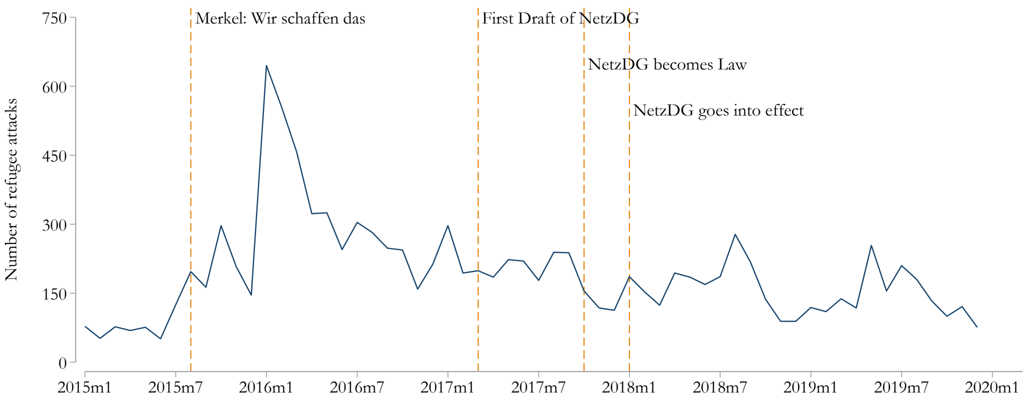
2 Background

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. This large inflow of asylum seekers 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.

In previous research, Müller and Schwarz (2021) have shown that this wave of anti-refugee crime coincided with an increase in anti-refugee sentiment online. In particular, 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 frequency of these hate crimes also drew the attention of the international news media (see for example New York Times, 2017).

In August 2015, Germany's Minister of Justice Heiko Mass demanded that social media companies should enforce existing laws prohibiting hate speech (Economist,

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 law.

2018).[3] Due to what he deemed insufficient action by the social media companies, Heiko 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.[4] 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 is "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 Euros if social media companies with more than 2 million registered users in Germany failed to remove hateful content within 24 hours of notice. The law also imposed an unprecedented transparency requirement for platforms to publish a biannual report on their content

[3] In an open letter, Mass wrote: "The internet is not a lawless space where racist abuse and illegal posts can be allowed to flourish ..."

[4] 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.

moderation activities (Heldt, 2019).

In Online Appendix A.1., we provide a theoretical framework that allows us to derive predictions about the first-stage 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 for 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.

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 offline hate crimes.

3 Data and Empirical Strategy

3.1 Data

We construct two separate data sets for our analysis. First, we build a panel of anti-refugee Tweets that allows us to study the first-stage impact of the NetzDG on the toxicity of online content. Given that one of the main concerns about online hate speech is that it may spill over into real-life action, our main analysis then focuses on anti-refugee hate crimes. Indeed, 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). For this reason, we use a municipality-quarter panel that allows us to analyze the impact of the NetzDG on anti-refugee incidents. We describe the main data sources for each panel in the following.

Tweet-Level Panel of Toxic Twitter Posts To provide evidence for the effects of the NetzDG on the toxicity of social media content, we create a Tweet-level panel measuring online toxicity based on Twitter data. We focus on data on 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 refugee crisis was largely responsible for the increase in online hate speech during this period, so we hypothesize that the NetzDG potentially changed the likelihood that refugee-related German Tweets contain hate speech. In total, this dataset contains 484,592 Tweets. We additionally scraped the followers of all major German parties. This dataset allows us to identify which Twitter users are following the AfD's Twitter account. We use Google's Perspective API (Wulczyn et al., 2017; Dixon et al., 2018) to obtain toxicity scores for each Tweet, which we use as a measure of hate speech. This API returns toxicity measures along the following five dimensions: toxicity, severe toxicity, identity attack, insult, profanity, and threat.

Appendix Table A.2 contains summary statistics for our sample of refugee Tweets. On average, refugee-related Tweets have a toxicity score equal to 0.41 and 5.6% 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 33% of Tweets in the sample were posted by AfD followers and 52% 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.

Municipality-Level Anti-refugee Incidents 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 were collected by the Amadeu Antonio Foundation and Pro Asyl (a pro-asylum NGO).[5] The anti-refugee hate crimes are categorized as arson of refugee homes, assault, incidents during protests, other cases of property damage (e.g., anti-refugee graffiti; henceforth "other" crimes), and suspected cases. The dataset also contains the date

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

and precise coordinates of each anti-refugee incident. We assign these incidents to municipalities using a shape files provided by the ©GeoBasis-DE/BKG 2016 website.[6]

Our data also includes measures of far-right Facebook usage and user activity from Müller and Schwarz (2019).[7] These data contain 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 data also contain information about the activity of each user and therefore allow us to construct the number of posts, likes, comments, and shares for each AfD user. 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.

To control for the number of Facebook users in a municipality, we create a simple measure based on Google searches. 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 a 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.

Finally, we also add municipality-level socio-economic controls and measures of voting and media consumption behavior. Data on other types of crimes by county and year come from the Bundeskriminalamt (BKA)'s Police Crime Statistics. Online Appendix A.2. provides a comprehensive overview of the data sources, variable definitions and summary statistics.

We visualize the main variation in Figure A.2. The map shows quintiles of AfD 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

[6] 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 size and population of these administrative areas.

[7] The underlying reproduction file is available here.

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.

3.2 Empirical Strategy

Our empirical analysis proceeds in two steps. First, we provide evidence that the NetzDG reduced the toxicity of online content. Second, we study the effect of the NetzDG on the frequency of anti-refugee hate crimes.

To investigate whether the NetzDG disproportionately reduced hateful online content, we estimate a canonical difference-in-differences regression of the following form:

$$Toxicity_{iut} = \theta \cdot AfD\ Foll._u \times Post\ NetzDG_t + \phi AfD\ Foll._u + \mu_t + \psi_{iut}, \qquad (1)$$

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 Twitter users who follow the AfD ($AfD\ Foll._u$) and the post-NetzDG dummy ($Post\ NetzDG_t$), which is equal to 1 starting in the fourth quarter of 2017 (October 1, 2017), when NetzDG took effect. Hence, our strategy compares the change in toxicity of refugee-related Tweets posted by AfD followers 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 AfD followers relative to other users, both mechanically (if Twitter removes toxic posts) and by deterring users from posting toxic content.

To measure the effect of the law on hate crimes, we exploit variation in the exposure of different German municipalities to online hate speech. Intuitively, we expect places with a high prevalence of this type of content to be disproportionately affected by the NetzDG relative to places with a low prevalence. 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 evolution in hate crimes.

This intuition gives rise to the following empirical strategy:

$$y_{it} = \theta \cdot AfD\ Users\ p.c._i \times Post\ NetzDG_t + \mathbf{X'_{it}}\beta + \gamma_i + \delta_t + \epsilon_{it}, \qquad (2)$$

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 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 overall change in the number of anti-refugee incidents (e.g., due to national news events). 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. The vector of control variables ($\mathbf{X_{it}}$) accounts for potential confounding variables (e.g., the municipality vote share of the AfD), which we also interact with the $Post\ NetzDG_t$ dummy. We cluster standard errors at the county level.[9]

4 Results

4.1 Did the NetzDG Reduce Online Toxicity?

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

Table A.3 presents the results from estimating equation (1). Columns (1) through (3) include all users who posted refugee-related Tweets. Columns (4) through (6) include only users who posted refugee-related Tweets and follow at least one of the major political parties (CDU/CSU, SPD, FDP, Green, Left, and AfD) on Twitter. The advantage of this restriction is that it is 1) more likely to capture German instead of German-speaking Twitter users and 2) more likely to capture a more homogeneous group of Twitter users with interest in politics. The different columns present specifications with varying sets of

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

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

fixed effects (e.g., user-specific linear time trends). In all specifications, the regressions indicate a significant reduction in the toxicity of Tweets by AfD followers relative to followers of other parties. The magnitude in column (1) suggests that the NetzDG was associated with a reduction in toxicity of around 8% relative to the mean.[10]

Figure 2 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. 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 2018q1. We also visualize the main results in Figure A.3. The bar graph visualizes the probability that a follower of a German party posts a Tweet with a toxicity above 0.8—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)—before and after the NetzDG. The graph shows two things. First, AfD followers are significantly more likely to produce highly toxic social media content. Second, the hatefulness of AfD followers' Tweets strongly decreases after the NetzDG, while the toxicity of Tweets sent by followers of other parties if anything slightly increases. Table A.4 presents a robustness exercise using the different measures of toxicity produced by Google's Perspective API. The effect is consistently significant across all toxicity measures except for the threat score.

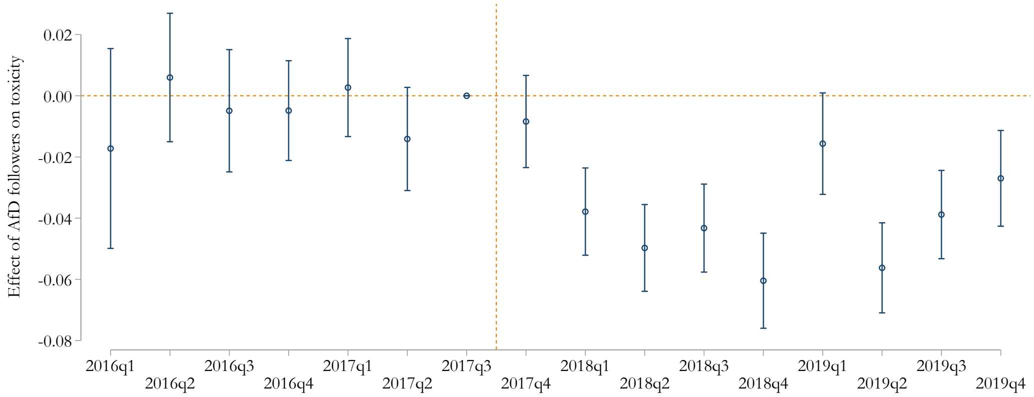
It is worth noting that we cannot disentangle whether these findings are driven by platforms deleting an increasing number of hateful Tweets after the implementation of the NetzDG or due to a deterrence effect leading users to self-censor. However, taken together, they do suggest that the NetzDG was 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.

4.2 Online Content Moderation and Hate Crimes

Baseline estimates. Table 1 shows our main results. Column (1) contains estimates of our baseline specification using Equation (2), controlling only for the log number of inhabitants to account for mechanical changes in hate crimes due to population differences.

[10] 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 2: Event Study Toxicity



Notes: This figure plots the coefficients from running an event study version of regression Equation (1). 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.

We then start controlling for some of the most relevant potential confounders. Column (2) adds Facebook users per capita in a municipality to account for a potential impact of social media channels that are not captured by the AfD usage measure. In the next two columns, we control for the access to broadband internet and the vote share of the AfD at the municipality level. Finally, in column (5) we include a wealth of additional control variables (see Appendix A.2. for details). 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 -0.008 point estimate indicates that a one standard deviation increase in AfD Facebook users per capita results in a -0.8 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 8% reduction in hateful online content we identified in the previous section.

One question is why we find that the effect on hate crime is already statistically significant in the 4th quarter of 2017 while the coefficient for the toxicity of tweets is negative but not yet statistically significant at conventional levels (p-value=0.27) before the first quarter of 2018. The most likely explanation for this finding is that Twitter was slower than Facebook to ramp up its content moderation efforts. An analysis by the European Union found that as of December 2017 Twitter removed 45% of hateful content that was reported to them (European Commission, 2019). In contrast, Facebook removed more than 80% in the same time frame. In particular in the German context, Facebook also faced considerably larger public scrutiny due to its larger user base; some commentators even referred to the NetzDG as the "Facebook Law" (see Spiegel, 2017).

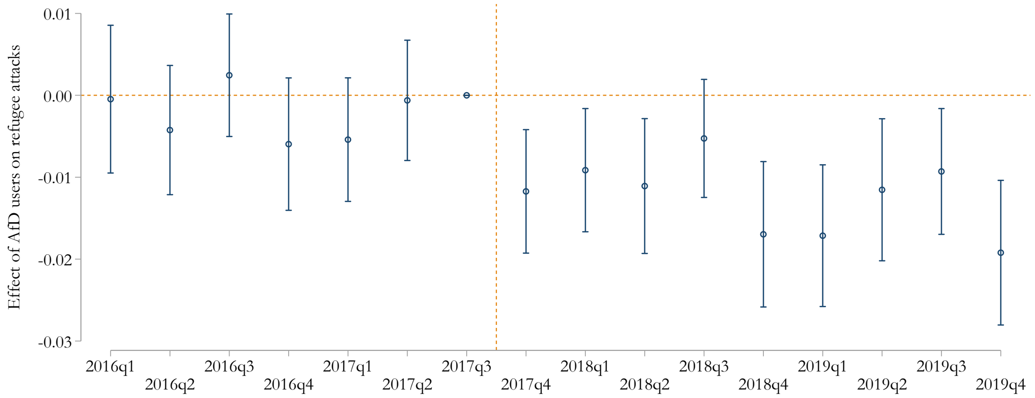
Table 1: Main Results

	Dep. var.: Asinh(Anti-Refugee Hate Crimes)				
	(1)	(2)	(3)	(4)	(5)
AfD Facebook users p.c. (std) × Post	-0.010***	-0.010***	-0.010***	-0.008***	-0.006***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)
Facebook users p.c × Post		0.004	0.004	0.004	0.004*
		(0.003)	(0.003)	(0.002)	(0.002)
Broadband internet × Post			0.001***	0.000	0.000
			(0.000)	(0.000)	(0.000)
AfD vote share × Post				-0.002***	0.004***
				(0.001)	(0.002)
Ln(Pop.) × Post	Yes	Yes	Yes	Yes	Yes
Municipality FE	Yes	Yes	Yes	Yes	Yes
Quarter FE	Yes	Yes	Yes	Yes	Yes
All Controls (19) × Post					Yes
Observations	71,456	71,456	71,456	71,008	68,736
Mean of DV	0.09	0.09	0.09	0.09	0.09
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 3 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 start of content moderation efforts in 2017q4. This negative effect appears to be persistent and stable over the two years following the NetzDG.

Figure 3: 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 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 90% 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 2, 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 2 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.

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.

Robustness. As a first robustness check, we provide a placebo test by exploring how the NetzDG impacted other categories of crimes that are unlikely to be motivated by online hate speech and should therefore be unaffected by the NetzDG. Table A.5 presents the results from a placebo test where we replace the dependent variable with the number of cyber crimes, property damages, robberies, thefts, and drug-related offenses. These data are only available on the county-year level, so we adjust our *Post* dummy to equal 1 for 2018 and after. We find that all coefficients are statistically insignificant, some are positive, and all are quantitatively small. While it can only be suggestive, these results provide some evidence that we are not just picking up a general reduction in crime in municipalities with many AfD users.

To probe the robustness of our findings, we perform four additional robustness checks. First, in Online Appendix Table A.6 we show that with the exception of arson there is an effect of the NetzDG is 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

Table 2: Heterogeneity by User Activity

	Dep. var.: Asinh(Anti-Refugee Hate Crimes)			
	(1)	(2)	(3)	(4)
AfD Facebook users p.c. (std) × Post	-0.016*** (0.005)	-0.016*** (0.005)	-0.016*** (0.005)	-0.016*** (0.005)
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				-0.005*** (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
Mean of DV	0.11	0.11	0.11	0.11
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. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

other property attacks. This makes it unlikely we are capturing changes in reporting of minor incidents.

Next, Online Appendix Table A.7 presents a battery of additional robustness exercises. In column (2), we show that our findings are robust to the inclusion of federal state × 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

West Germany, or municipalities without anti-refugee incidents, without AfD users, or with few refugees per capita (columns (4), (5), (6), and (7), respectively). Throughout these exercises, our results remain highly statistically significant.

Third, Table A.8 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 for 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 non-parametric specification does not rely on any functional form assumptions and simply picks up changes in the mean number of anti-refugee incidents after the NetzDG in a canonical difference-in-differences setting. Lastly, this transformation also rules out that our findings could be driven by non-homogenous treatment effects in our two-way fixed effects estimation (De Chaisemartin and D'Haultfoeuille, 2022), as our results also hold in this dummy specification.

Finally, Table A.9 shows that our estimates remain statistically significant irrespective of the level of clustering of the standard errors. More specifically, we show our main results clustered at: 1) the county level (baseline), 2) the county and quarter level, 3) the municipality level, or 4) the municipality and quarter level. Neither of these levels makes any difference for our findings.

5 Discussion

Much attention has been devoted to the spread of hateful content on social media. The controversial 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 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

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.

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.

European Commission (2019). Code of Conduct on countering illegal hate speech online. *https://ec.europa.eu/info/sites/default/files/code_* *of_ conduct_ factsheet_ 7_ web. pdf*.

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

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 (2019). From Hashtag to Hate Crime: Twitter and Anti-Minority Sentiment. *Working Paper*.

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.

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

Spiegel (2017). Maas rüstet personell gegen Facebook & Co. auf. *https://www.spiegel.de/netzwelt/netzpolitik/heiko-maas-ruestet-personell-gegen-facebook-co-auf-a-1170588.html*.

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.[11] 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; $d\tau > 0$.[12] 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; $dT^H/d\tau < 0$.[13]

[11] 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.

[12] 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 euros for violating the NetzDG law (Bundesamt für Justiz, 2019).

[13] 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

The municipality-level panel dataset we construct is based on the replication data from Müller and Schwarz (2021), available from the journal's website. We briefly describe each type of data we use below and refer the reader to Müller and Schwarz (2021) for additional details.

Anti-Refugee Incidents. The source of these data is the Antonio Amadeu Foundation and Pro Asyl. For the time period from January 2016 to December 2019, all 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. Figure A.2 shows the location of the anti-refugee incidents in our observation period for each German municipality.

Municipal-Level Facebook Measures We construct a measure of exposure to right-wing populist social media at the municipal level using information from Facebook. To the best of our knowledge, there is no municipality-level dataset on Facebook usage in Germany. To construct our measures, we hand-collect user location data by using the unique user identifiers provided by the Facebook Graph API. Due to Facebook's privacy policy, we are only able to collect this information for people who make it publicly available.

We are interested in each municipality's exposure to hateful content targeting refugees on social media. We proxy for this exposure based on the locations of Facebook users active on the AfD page. In total, we can identify 93,806 users who interacted with the page at least once.[14] We can identify the place of residence for 34,389 of these users, covering at least one user for 3,563 of the 4,466 municipalities in our dataset. Figure A.2 plots the geographical distribution of AfD users per capita. We also construct measures of the intensity of AfD user activity in a municipality based on the frequency of posts, likes, comments, and shares.[15]

[14] The Facebook API does not provide data on which users "like" a page but only on users who *interact* with a page, e.g. by liking another user's comment. As a result, the total number of user IDs we have is smaller than the more than 300,000 people who had liked the AfD Facebook page as of 2017.

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

A.2.1 Auxiliary and Control Variables

The control variables we use come from several sources. 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 Gesellschaft der deutschen Zeitungen (ZMG)" (Society for Market Research of German Newspapers), which we normalize using a municipality's population.

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	0.03	0.00	0.03	0.80	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
Log(Facebook User per capita)	0.10	0.41	0.00	0.05	12.55	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

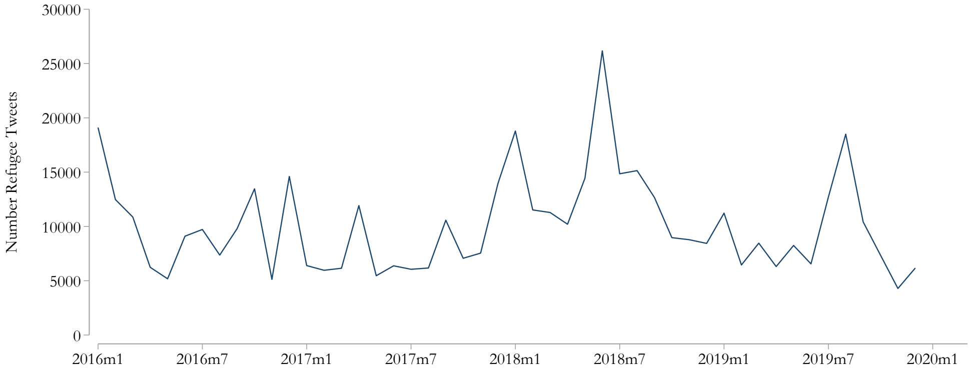
Notes: This table displays the mean, standard deviation, median, minimum, maximum, and number of observations of our main outcome, main variables of interest, and controls.

Table A.2: Summary Statistics Toxicity Refugee Tweets

Variable	Mean	SD	p50	Min	Max	N
Toxicity Measures						
Toxicity	0.41	0.22	0.00	0.42	1.00	484,592
Sev. Toxicity	0.32	0.24	0.00	0.29	1.00	484,592
Identity Attack	0.53	0.25	0.00	0.52	1.00	484,592
Insult	0.35	0.21	0.00	0.35	1.00	484,592
Profanity	0.23	0.21	0.00	0.12	1.00	484,592
Threat	0.41	0.29	0.00	0.27	1.00	484,592
User Measures						
AfD Twitter Followers	0.33	0.47	0.00	0.00	1.00	484,592
Party Twitter Followers	0.52	0.50	0.00	1.00	1.00	484,592

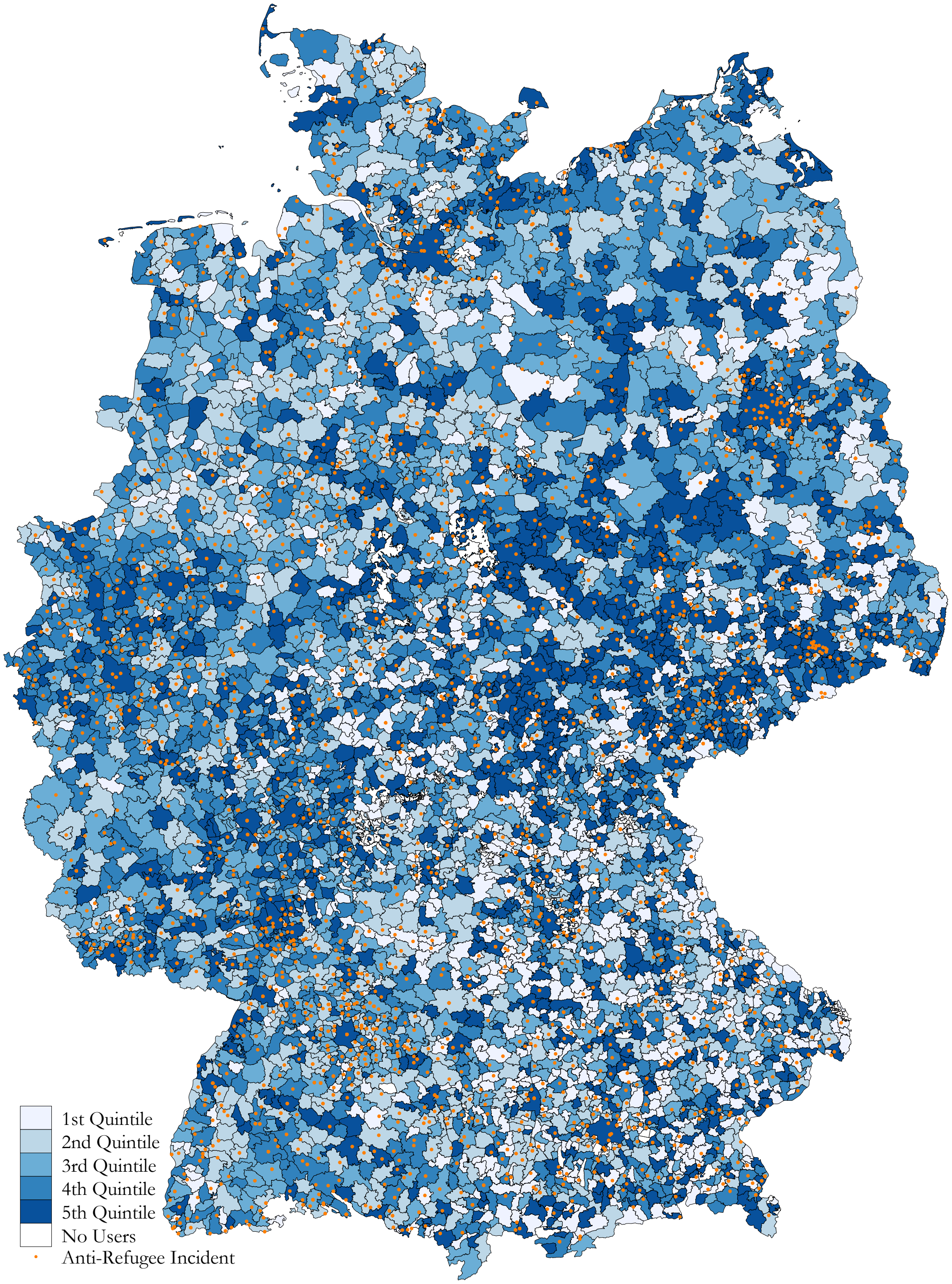
Notes: This table displays the mean, standard deviation, median, minimum, maximum, and number of observations for the main variables which are in the analysis of the toxicity of refugee Tweets.

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: Map AfD Facebook Users and Anti-Refugee Incidents



Notes: The shading of the maps indicate the quintiles of the distribution of AfD users per capita for the municipalities in Germany. Each orange dot indicates an anti-refugee incident in our data.

A.3. Additional Results

A.3.1 Additional Results for the Toxicity of Tweets

Table A.3: Regressions Toxicity

	Dep. var.: Toxicity Measures					
	All Twitter Users			Users Following Any Party		
	(1)	(2)	(3)	(4)	(5)	(6)
AfD followers × Post	-0.032*** (0.004)	-0.016*** (0.002)	-0.017*** (0.004)	-0.036*** (0.003)	-0.021*** (0.003)	-0.021*** (0.005)
AfD Follower FE	Yes			Yes		
User FE		Yes	Yes		Yes	Yes
Day FE	Yes	Yes	Yes	Yes	Yes	Yes
Linear Time Trend			Yes			Yes
Observations	484,592	433,499	433,499	250,550	237,964	237,964
Mean of DV	0.41	0.42	0.42	0.42	0.42	0.42
R^2	0.14	0.29	0.36	0.14	0.25	0.31

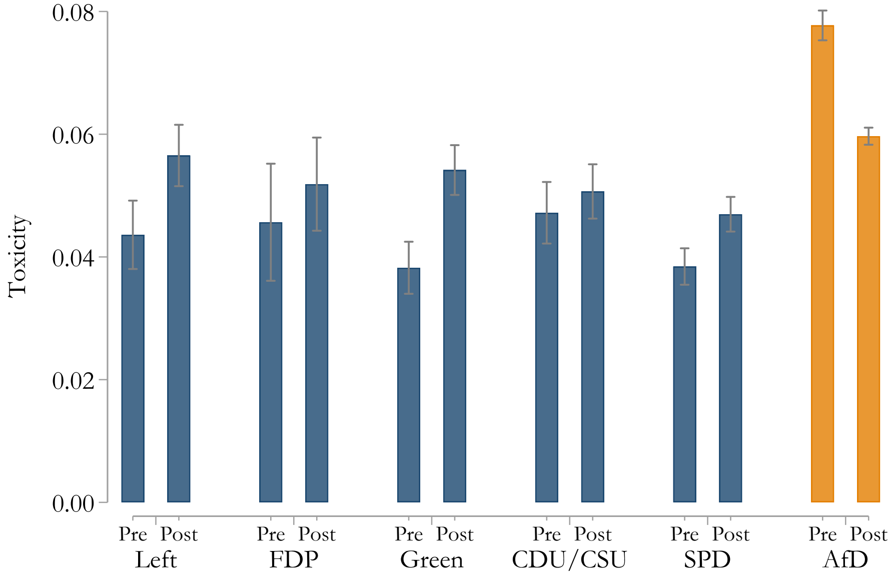
Notes: This table presents the results of estimating Equation (1) where the dependent variable is the toxicity of Tweets containing the word ”Flüchtling” (refugee) (bounded between 0 and 1). *AfD followers* is an indicator variable that is equal to 1 if a Twitter user follows the AfD's account. All regressions control for either AfD follower or user fixed effects, as well as day fixed effects. In column (3) and (6), we additionally include user-specific linear time trends. The first three columns are estimated for the sample of all Twitter users who posted at least one refugee Tweet. Columns (4-6) restrict the sample to users who follow at least one major German party on Twitter. Robust standard errors in parentheses are clustered by user. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.4: Robustness: Toxicity Measures

	Dep. var.: Measure of Toxicity					
	Toxicity	Severe Toxicity	Identity Attack	Insult	Profanity	Threat
	(1)	(2)	(3)	(4)	(5)	(6)
AfD followers × Post	-0.032*** (0.004)	-0.034*** (0.004)	-0.040*** (0.004)	-0.034*** (0.004)	-0.033*** (0.005)	0.004 (0.008)
AfD Follower FE	Yes	Yes	Yes	Yes	Yes	Yes
Day FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	484,592	484,592	484,592	484,592	484,592	484,592
Mean of DV	0.41	0.32	0.53	0.35	0.23	0.41
R^2	0.14	0.14	0.13	0.12	0.12	0.16

Notes: This table presents the results of estimating Equation (1) where the dependent variable is the measure of toxicity in the top, bounded between 0 and 1, calculated based on Tweets containing the word refugee ("Flüchtling"). *AfD followers* 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 user. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Figure A.3: Frequency of Highly Toxic Tweets by Party



Notes: These bar graphs show the frequency of Tweets with a toxicity larger than 0.8 depending on which German party users follow. We differentiate between the period before and after the NetzDG, where observations starting in 2017q4 (when the law was enacted) are classified post-NetzDG.

A.3.2 Additional Results for Hate Crimes

Table A.5: Effect on Other Crimes

	Dep. var.: Asinh(Crime)				
	Cyber	Prop. Damage	Robbery	Theft	Drug
	(1)	(2)	(3)	(4)	(5)
AfD Facebook users p.c. (std) × Post	-0.023	-0.009	0.018	-0.002	0.012
	(0.019)	(0.007)	(0.012)	(0.004)	(0.009)
Ln(Pop.) × Post	Yes	Yes	Yes	Yes	Yes
County FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Observations	2,005	2,005	2,005	2,005	2,005
Mean of DV	5.47	7.71	4.82	9.55	6.99
R^2	0.90	0.98	0.96	1.00	0.96

Notes: This table presents the results from estimating a county-year-level version of Equation (2). The dependent variable is the inverse hyperbolic sine of the number of crimes in the category indicated at the top of the table in a given county and year. *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 control for county, year fixed effects and the logarithm of population interacted with *Post*. Robust standard errors in parentheses are clustered by county. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.6: 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	-0.008***	-0.000	-0.003***	-0.001**	-0.006***	-0.001**
	(0.003)	(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
Mean of DV	0.09	0.00	0.02	0.00	0.07	0.00
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 anti-refugee hate crimes of a specific type (indicated at the top). *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. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.7: Robustness

	Dep. var.: Asinh(Anti-Refugee Hate Crimes)						
		State	Exclude	Exclude	Exclude	Exclude	Exclude
	Baseline	× Quarter FE	Q1 2016	West Germany	Attack= 0	AfD User= 0	Few Refugees
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
AfD Facebook users p.c. (std) × Post	-0.008***	-0.007***	-0.008***	-0.016***	-0.019***	-0.010***	-0.013***
	(0.003)	(0.002)	(0.002)	(0.005)	(0.006)	(0.003)	(0.004)
Ln(Pop/) × Post	Yes	Yes	Yes	Yes	Yes	Yes	Yes
AfD vote share × Post	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Facebook users p.c × Post	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Broadband internet × Post	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Quarter FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fed. State × Quarter FE		Yes					
Observations	71,008	71,008	66,570	16,272	36,384	64,736	56,656
Mean of DV	0.09	0.09	0.08	0.15	0.17	0.09	0.11
R^2	0.44	0.45	0.45	0.51	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 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. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.8: Robustness: Specification

	Dep. var.:					
	Asinh	Count	Ln(p.c.)	Asinh	Count	Ln(p.c.)
	AfD User per Capita			High AfD Usage Dummy		
	(1)	(2)	(3)	(4)	(5)	(6)
AfD Facebook users p.c. (std) × Post	-0.008***	-0.022***	-0.006***			
	(0.003)	(0.007)	(0.002)			
High AfD Usage × Post				-0.025***	-0.081***	-0.020***
				(0.007)	(0.024)	(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
Mean of DV	0.09	0.14	-9.09	0.09	0.14	-9.09
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.9: Robustness: Standard Errors

	Standard Errors Clustered by:			
	County	County & Quarter	Municipality	Municipality & Quarter
	(1)	(2)	(3)	(4)
AfD Facebook users p.c. (std) × Post	-0.008*** (0.003)	-0.008*** (0.002)	-0.008*** (0.002)	-0.008*** (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
Mean of DV	0.09	0.09	0.09	0.09
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 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. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.