University of Zurich

Department of Economics

Working Paper Series

ISSN 1664-7041 (print)
ISSN 1664-705X (online)

Working Paper No. 446

Censorship in Democracy

Marcel Caesmann, Janis Goldzycher, Matteo Grigoletto and Lorenz Gschwent

June 2024

Censorship in Democracy

Marcel Caesmann[*], Janis Goldzycher[†], Matteo Grigoletto[‡], Lorenz Gschwent[§]

June 4, 2024

Abstract

The spread of propaganda, misinformation, and biased narratives from autocratic regimes, especially on social media, is a growing concern in many democracies. Can censorship be an effective tool to curb the spread of such slanted narratives? In this paper, we study the European Union's ban on Russian state-led news outlets after the 2022 Russian invasion of Ukraine. We analyze 775,616 tweets from 133,276 users on Twitter/X, employing a difference-in-differences strategy. We show that the ban reduced pro-Russian slant among users who had previously directly interacted with banned outlets. The impact is most pronounced among users with the highest pre-ban slant levels. However, this effect was short-lived, with slant returning to its pre-ban levels within two weeks post-enforcement. Additionally, we find a detectable albeit less pronounced indirect effect on users who had not directly interacted with the outlets before the ban. We provide evidence that other suppliers of propaganda may have actively sought to mitigate the ban's influence by intensifying their activity, effectively counteracting the persistence and reach of the ban.

Keywords: Censorship, Policy effectiveness, Text-as-data, Media slant
JEL Classification: D72, D78, L82, P16

Acknowledgments:
We are grateful to Joop Adema, Elliot Ash, Kai Gehring, Carlo Schwarz, Nikita Zakharov and seminar participants at the University of Zurich, RTG Regional Disparities and Economic Policy, University of Duisburg, as well as participants and discussants at the CINCH Research Agenda for Ukraine workshop, 5$^{\text{th}}$ Monash-Warwick-Zurich Text-as-Data workshop, Annual Meeting of the Austrian Economic Association (NOeG) 2023, Digital Democracy Workshop DigDemLab Zurich, 17$^{\text{th}}$ CESifo Workshop on Political Economy, 2$^{\text{nd}}$ CESifo/ifo Junior Workshop on Big Data, 7$^{\text{th}}$ Conference on Political Economy of Democracy and Dictatorship, 17$^{\text{th}}$ RGS Doctoral Conference, European Public Choice Society 2024, FRIAS conference Understanding the Rise of Autocrats in the 21st Century, for their helpful comments.

[*]University of Zurich, e-mail: *marcel.caesmann@econ.uzh.ch*
[†]University of Zurich, e-mail: *goldzycher@cl.uzh.ch*
[‡]University of Bern, Wyss Academy for Nature at the University of Bern, e-mail: *matteo.grigoletto@unibe.ch*
[§]University of Duisburg-Essen, RTG Regional Disparities and Economic Policy, e-mail: *lorenz.gschwent@uni-due.de*

1 Introduction

Slanted media can be very effective in swaying opinions and behavior (Enikolopov, Petrova, and Zhuravskaya 2011; Yanagizawa-Drott 2014; Adena et al. 2015). Recognizing this, autocratic regimes extensively use censorship of media to suppress dissent and manage the flow of information within their borders, while employing slanted narratives to exert influence abroad (Guriev and Treisman 2022). Concerns regarding the spread of misinformation, foreign propaganda and biased narratives, especially on social media, have been rising in recent years in democratic societies, in particular since the Russian interference in the 2016 US presidential election.[1] This pivotal event sparked heightened attention on information manipulation threats, such as the weaponization of information as part of Russia's 'asymmetric warfare' strategy in the context of its invasion of Ukraine or Chinese influence over TikTok, leading to the PAFACA act passed by US congress in 2024 to counteract foreign media influence (US 118th Congress, 2024). Given the trade-off between using censorship and upholding the conviction that free speech and independent media are the backbones of the democratic order itself, it is crucial to understand the effectiveness and consequences of media regulation in the context of liberal democracies. While there is some evidence on the effects of censorship in authoritarian regimes (Chen and Yang 2019; Becker, Pino, and Vidal-Robert 2021), systematic evidence on the effectiveness and consequences of government-imposed censorship in a democratic context is very limited.

To shed light on the effects of censorship in democracies, we study a ban on Russian state-backed media outlets implemented by the European Union in March 2022 in the context of Russia's invasion of Ukraine. The ban aimed to counteract the spread of Russian narratives provided by seemingly independent news sources to justify the invasion and influence public opinion.[2] The unprecedented decision to completely ban all activity of the two most prominent of these outlets – Russia Today and Sputnik – was taken virtually overnight. It affected all their channels, including online platforms, in the European Union from the 2^{nd} of March onwards. We focus on the social media platform Twitter/X, a pivotal platform in shaping public opinion and fueling both offline and online political activity (Allcott and Gentzkow 2017; Acemoglu, Hassan, and Tahoun 2018). Twitter serves as a platform where narratives – often led by misinformation or radical influential users – can become extreme (Müller and Schwarz 2023). Its influence extends beyond digital boundaries, with narratives and stories that gain traction on the platform frequently making their way into mainstream and traditional media (Cage, Herve, and Mazoyer 2020).

To measure pro-Russian or pro-Ukrainian media slant, we build on Gentzkow and Shapiro (2010) and Gennaro and Ash (2023), establishing two distinct poles of slant: one favoring Russia and the other Ukraine. The slant of a tweet is determined by its proximity ratio to these two poles and captures the nuance of discussion in the context of Twitter. Positive numbers indicate pro-Russian while negative

[1] For analyses of the event, see for example work by Badawy et al. (2019) and Eady et al. (2023).

[2] As highlighted by the European Court of Justice's upholding, the ban was implemented on the grounds of the prohibition of "propaganda for war" in international law (International Covenant on Civil and Political Rights, UN 1966).

numbers refer to pro-Ukrainian slant.[3]

To assess the effects of censorship in democratic contexts, we exploit geographic and temporal variation in the implementation of the ban on Russian state-controlled media, applying a difference-in-differences approach. Our analysis compares social media discourse in EU countries enforcing the ban – Austria, France, Germany, Ireland, and Italy – with that in non-EU countries – Switzerland and the United Kingdom – where no restrictions on Russia Today or Sputnik were applied during our study period. We analyze 775,616 English tweets from 133,276 users on the war in Ukraine. We conceptualize the ban as a supply shock and study its effect from three analytical angles. First, we focus on the intensive margin, i.e. the degree of slant used in tweets. Second, we study the extensive margin, i.e. the volume of slanted content created in the form of tweets. Third, we investigate the spread of slanted content through retweets. By examining these three aspects – extremism, production, and spread – we aim to comprehensively understand the effects of the ban. Based on this conceptual and analytical framework, we proceed in three steps to investigate how the ban affected users directly and indirectly.

First, we examine the impact of the ban on a specific subset of users who, at any time before the ban, had interacted at least once with the banned outlets. We refer to this group as *interaction users*. Despite their relatively small number in our sample, these users are highly active and make up for a significant fraction of tweets on the conflict. They are more likely to support Russian propaganda and serve as a crucial link in the network, potentially bridging extreme views and misinformation from government propaganda outlets to the broader user base. Considering the aim of the ban in decreasing the spread of narratives originating from state-led Russia Today and Sputnik, these users represent a prime sample of interest to assess the effect of the ban.

We find that the ban reduced the average slant of the *interaction users*, 63.1% compared to the pre-ban mean. In turn, we observe no significant change in the proportion of tweets and retweets that are pro-Russia – defined as tweets with more than one standard deviation from the neutral point towards the Russian pole. Exploring the heterogeneity of effects, we find that effects are most pronounced among *interaction users* who used the highest pro-Russian slant before the ban on average. Specifically, users in the top 25% of the pre-ban average slant distribution experienced a reduction of 0.15 standard deviations in their slant compared to their counterparts in non-EU countries while there is no clear effect for the bottom 75% of the distribution.

Second, we investigate the temporal impact observed among the *interaction users*. Despite detecting a significant effect on our slant measure in the pooled pre-post comparison, results using an event-study design on a daily level suggest that the influence of the ban faded within just a few days. This pattern is confirmed when we detect a meaningful and significant effect on the average slant in the first week after the ban, while effects in the second week are smaller in magnitude and with confidence intervals including

[3] Throughout this paper, references to a "pro-Russia" or "pro-Russian" slant specifically denote content that aligns with the Russian Government's perspectives at the time of the Ukraine invasion. This terminology is not intended to generalize or suggest that the Russian government's actions or policies reflect the opinions of the entire Russian population (analogously for "pro-Ukrainian" slant).

zero.[4] Overall, the results suggest an immediate but short-lived effect on the users who interacted with Russia Today and Sputnik before the ban.

Next, we study the indirect effects of the ban on users who did not directly interact with the banned outlets. We label these individuals as *non-interaction users*. We assess the ban's impact on these users across the same metrics of intensive and extensive margins as we did for the *interaction users*. We find that the ban reduced pro-Russian slant among the *non-interaction users*, despite to a lesser degree, resulting in a decrease of approximately 17.3% from pre-ban levels of slant, in contrast to the 63.1% observed among *interaction users*. Notably, we do find effects on the extensive margin for this group of users, specifically in a reduction of the share of pro-Russian retweets. This suggests that *non-interaction users* are deprived of slanted content they are able and willing to share.

Our results show that the ban had an effect, particularly on those users who interacted with the banned outlets before the implementation. However, this effect fades quickly. In the third step of our analysis, we investigate the mechanisms that might have compensated for the ban's effect, effectively re-balancing the supply of pro-Russia slanted content. This part of our study specifically examines users identified as *suppliers* of slanted content. We define as *supplier* any user who, during a given time period, produced or shared any content that was more than one standard deviation towards the pro-Russia pole.

The first potential mechanism we examine is the entry of new *suppliers* following the ban. We assess the proportion of users identified as *suppliers* in both EU and non-EU countries, before and after the ban was implemented. Additionally, we explore a potential rise in the number of pro-Russian bots. In both analyses, we find that although there is an increase in the number of *suppliers* within the EU post-ban this increase is actually outpaced by the rise in *suppliers* observed in non-EU regions. Thus, while the augmented number of *suppliers* might contribute to mitigate the impact of the ban, it seems unfit to explain the different trajectories in EU and non-EU countries.

The second mechanism we investigate concerns the behavior of users who were already disseminating pro-Russia content prior to the ban. It is plausible that these individuals were either competing with or aiding the banned outlets in spreading slanted content. With the ban in place, they might have moved to occupy the void created. Our analysis of the ban's impact on these users yields several key findings. Overall, it appears that European *suppliers* affected by the ban reduced both their content's slant and their share of pro-Russian retweets compared to their non-EU counterparts. When examining the heterogeneity of this effect based on pre-ban activity levels, it becomes evident that the primary contributors to this effect were *suppliers* who were moderately active before the ban. In contrast, the most active *suppliers* seemed unaffected by the ban. Moreover, an analysis of the very top *suppliers* in the activity distribution prior to the ban suggests – with caution due to the small sample size – that the most active *suppliers* may have even increased the production of new pro-Russian content in response to the ban.

This paper contributes to several strands of the literature. First, we contribute to the literature on censorship. Conceptual works by Shadmehr and Bernhardt (2015) and Gehlbach and Sonin (2014) highlight a trade-off between censoring and allowing unbiased information that autocratic rulers face.

[4] We limit the analysis to two weeks after the ban, as in the third week post-ban, the UK adopted a ban on the same outlets as well depriving us of a sizeable control group. Further, in the course of this project the access to Twitter/X API for researchers was restricted and limited the option to collect additional data.

Censoring comes with the cost of signaling to the population the attempt to control the information and readers might decide to take distance from the news outlets if they do not meet their need for informational content. Empirical evidence on the effects of censorship is limited. In autocratic contexts, Chen and Yang (2019) study the impact of providing citizens with access to uncensored internet and find effects on beliefs, attitudes, and intended behavior. Becker, Pino, and Vidal-Robert (2021) show the impact of censorship imposed by the Catholic Church during the Counter-Reformation preventing the diffusion of Protestant material. In the context of Russia, Simonov and Rao (2022) show how outlet-specific characteristics attract readers to government-controlled media and how readers, once there, do not change information sources. Overall, the existing evidence suggests that censorship can be effective in impeding the spread of information and changing citizens' beliefs and behaviors in autocratic contexts.

In democratic contexts, there is even more limited evidence of the effectiveness of censorship. Bjørnskov and Voigt (2021) study the effect of constitutional provisions in preventing media censorship in the aftermath of terrorist attacks. This is one of the few examples investigating the interaction between functioning constitutional systems and censorship of media, in line also with the work by Kellam and Stein (2016), providing evidence that powerful presidents can be threatening to media freedom even in democratic contexts. There is growing attention to weaker forms of self or state-mandated regulation of media platforms. Information withholding that is only targeting certain outlets is referred to as "selective censorship" (Guriev and Treisman 2022). So far, empirical analyses of this form of censorship have been rare. Corduneanu-Huci and Hamilton (2022) find in a cross-country analysis that media outlets that likely reach the median voter have a higher chance of being censored in both autocracies and democracies. They further argue that censors undertake a cost-benefit analysis weighing an outlet's audience's perceived political danger against legal and reputation costs. In the case of the EU's ban on Russia Today and Sputnik, judiciary viability needs to be particularly considered. Baade (2022) argued that the ban can be viably justified based on the UN's prohibition of "propaganda for war",[5] but this argument could not be readily applied to different outlets. The European Court of Justice's upholding of the ban against a complaint by Russia Today in July 2022 indeed relied on this prohibition.[6] Therefore, studying this ban provides a unique natural experiment for studying the causal effects of censorship in a democratic context.

Second, we also add to the rich literature investigating the political economy of social media (see Campante, Durante, and Tesei (2023) for an overview). The effects of the advent of social media on political systems are not clear yet. On one side, social media appears to facilitate the spread of populism (Campante, Durante, and Sobbrio 2018; Guriev, Melnikov, and Zhuravskaya 2021), xenophobia (Bursztyn et al. 2019), boosts a trend towards political polarization (Halberstam and Knight 2016; Levy 2021; Müller and Schwarz 2023) and reduce subjective well-being (Allcott et al. 2020). On the other side, some scholars argue not only that segregation offline is stronger than that online (Gentzkow and Shapiro 2011), but also that social media could play a role in decreasing polarization (Barbera 2014).

An emerging literature in this space is concerned with the effect of online content moderation (Chandrasekharan et al. 2017; Jhaver et al. 2021; Jiménez-Durán 2023). Morales (2020) studies the effect of

[5] See Article 20 of the International Covenant on Civil and Political Rights.
[6] For more information, see the official statement by the European Court of Justice on the ruling.

banning bots programmed to retweet the Venezuelan president Nicolás Maduro's tweets, showing that this makes the discussion on Twitter become more critical of the president. Ershov and Morales (2021) find that policies/nudges around the 2020 US election aimed at decreasing the sharing of misinformation led to a stronger decrease in the sharing of more factual news. Closest to our paper are studies by Müller and Schwarz (2022) and Jiménez Durán, Müller, and Schwarz (2022). They study the effect of banning Trump's account on reducing toxicity amongst his followers and the effects of a German regulation on removing online hate speech directed towards refugees. In line with our results, both studies show that online content moderation can curb toxicity and hate speech online. We go beyond the existing evidence on two dimensions. First, our study can exploit cross-country variation in who is affected by the ban with a more natural control group to identify effects. Second, we study the response of users – the pre-ban *suppliers* of media slant – in filling the gap left by banned outlets.

Third, our study contributes to the understanding of media slant, which can be interpreted as an indicator of state-led narratives, often associated with state propaganda. This expands the literature on the effects of propaganda (Enikolopov, Petrova, and Zhuravskaya 2011; Yanagizawa-Drott 2014; Adena et al. 2015), which traditionally examines the impact of increasing propaganda exposure. Our research setting allows us to explore the consequences of reducing exposure to media slant, providing a new dimension to the policy debate on media regulation. Moreover, we introduce a novel, data-driven approach to measuring media slant, leveraging advances in text-as-data methodology. This new measure, inspired by the works of Gentzkow and Shapiro (2011) and Gennaro and Ash (2023), offers a more nuanced understanding of media slant and its effects.

The remainder of the paper is structured as follows. Section 2 discusses the data used in our analysis while section 3 provides an overview of our measure of media slant and the approach used to obtain the measurement. Section 4 gives an overview of the framework and methods used for the analysis. Finally, we present our results on the effect of censorship in democracy in section 5 and provide a concluding discussion in section 6.

2 Data

In this paper, we use two main samples of tweets. The first one consists of more than 15,000 tweets by official Ukrainian and Russian government accounts, collectively labelled as *government tweets* (GT), which will be further discussed in Section 3. The second sample constitutes the main dataset for our analysis, to which we refer from now on as to the *users' tweets* (UT). This dataset covers the general European Twitter discussion about the Russo-Ukrainian conflict between February 19^{th}, 2022, and March 15^{th}, 2022. We provide detailed information about the extraction of this dataset in the Appendix A.

For the *users' tweets* – due to limitations in accessing the Twitter API[7] – we restrict our extraction to the following European countries: Austria, France, Germany, Ireland, Italy, Switzerland, and the United Kingdom. We also focus exclusively on tweets in the English language for several reasons. First, the most important branches of the Russian state-led propaganda outlets were those operating in English. Second,

[7] Free API access for researchers was unfortunately suspended shortly after Elon Musk's takeover of Twitter in October 2022 and during this research project.

considering the international reach of the Russo-Ukrainian conflict, much of the discussion on Twitter was happening in English. Third, for the *government tweets*, we selected officials posting extensive English content to signal their willingness to spread their narrative beyond the boundaries of their countries.

The resulting sample consists of 775,616 English-language tweets by 133,276 users. We show descriptive statistics on the tweet level in Table 1, Panel A. 5.7% of the tweets in our sample can be classified as being pro-Russian slant, while 10% are a retweet of pro-Russian slant. The number of own tweets and retweets is fairly balanced, with the latter representing 53% of the sample. Notably, tweets of our sample are retweeted extensively, with the median being 26 retweets per tweet. We show descriptive statistics on the user level in Table 1, Panel B. On average, each user of our sample produces 2.7 tweets and 3.1 retweets covered in our sample. Of all users, 3.7% had direct contact with the banned outlets before the ban, as in a reply or a retweet of the outlets' accounts. In the Appendix Table E.1, we provide the same summary statistics after cleaning our sample of 2,489 potential bots. In Appendix F Table F.1, we also show the descriptive statistics after dropping 389 accounts in our sample that were created only after the ban was enforced.

In Figure 1, we delve into the geographical distribution of tweets in our dataset, mapping each tweet from the UT sample, according to the location of the user producing the tweet. As expected, a significant portion originates from the United Kingdom. This is attributable to the country's substantial Twitter user base and our emphasis on English-language tweets. While the UK's representation is pronounced, it is important to keep in mind the two areas that we compare in our study encompass the UK and Switzerland on the one side – as countries where the ban was not applied – and Austria, France, Germany, Ireland, and Italy on the other side – as countries where the ban was applied.

3 Measuring pro-Russian and pro-Ukrainian slant

In the context of the Russian invasion of Ukraine in 2022, we conceptualize the discourse on the war as defined by a one-dimensional continuum between two narrative poles: pro-Russia and pro-Ukraine. In our analysis, discussions about the conflict occupy positions along this spectrum, with a tweet's content being closer to one pole or the other, indicating its narrative slant. This proximity to either pole reflects the intensity of its alignment, with content equidistant from both poles representing a neutral stance, thus supporting neither side strongly. By adopting this framework, we emphasize the importance of capturing the nuanced spectrum of discussions, arguing that it would be overly simplistic and inaccurate to categorize tweets exclusively as pro-Russia or not, hence acknowledging the complexity of public discourse and opinion formation during the conflict.

To obtain a quantifiable and tractable measure of pro-Russian and pro-Ukrainian media slant, we adopt a procedure proposed by Gennaro and Ash (2023), drawing inspiration from earlier work on media slant by Gentzkow and Shapiro (2011). This approach is both simple and powerful, relying on the measurement of language similarity in tweets by European users discussing the war relative to two distinct ideological poles. More specifically, we calculate the cosine similarity of a tweet's language to what we define as the 'pro-Russian pole' and compare this to its similarity to what we define as the 'pro-Ukrainian

Table 1: *Summary statistics*

Panel A: Tweet level

	Mean	Median	St. Dev.	Min.	Max.
Dependent Variables					
Propaganda ratio	-.00011	.042	1	-4	4.8926959038
Russian propaganda tweet	.057	0	.23	0	1
Russian propaganda retweet	.1	0	.3	0	1
Tweet type					
Retweet	.53	1	.5	0	1
No. of words	25	23	11	1	108
No. of mentions	1.6	1	2.4	0	50
No. of hashtags	.44	0	1.6	0	42
No. of Observations	775,616				

Panel B: User level

	Mean	Median	St. Dev.	Min.	Max.
User behavior					
No. tweets from user	2.7	1	12	0	1,528
No. retweets from user	3.1	1	11	0	616
No. replies from user	.52	0	2.1	0	202
No. russian propaganda tweets	.33	0	1.6	0	300
No. russian propaganda retweets	.6	0	2.3	0	138
Interacted with RT/Spk	.037	0	.19	0	1
No. retweets of RT/Spk	.001	0	.044	0	6
Region					
European Union	.39	0	.49	0	1
No. of Observations	133,276				

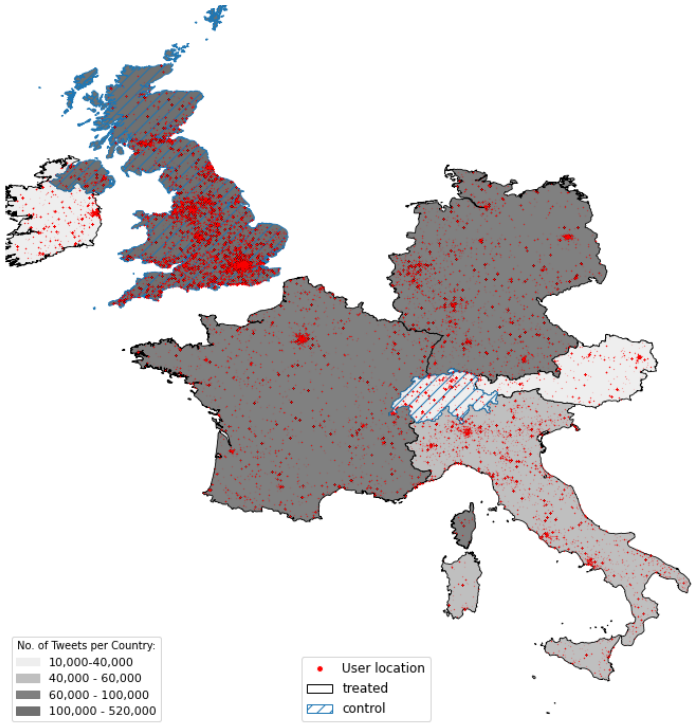
Notes: Panel A reports descriptive statistics for the sample of tweets used in the analysis, posted by users that we could locate in the countries of interest: Austria, France, Germany, Ireland, Italy, Switzerland and the United Kingdom. Tweets were extracted using the Historical Twitter APIv2, with the query: *ukrain* OR russ* OR NATO OR OTAN*, in the time window between February 19^{th} and March 15^{th} 2022. Panel B reports descriptive statistics on user characteristics, for users that posted tweets used in our analysis and described in Panel A. In both panels, for all variables we report mean, median, standard deviation, minimum, and maximum values. Appendix Tables E.1, F.1 and G.1 show the same descriptive statistics excluding users that are plausible bots, excluding accounts created only after the ban and using 0, instead of 1, threshold to define the binary variables of pro-Russia tweets and retweets, respectively.

pole'. The critical decision in this method lies in determining the content that constitutes each pole.

We construct our ideological poles using content disseminated on Twitter by key figures within the Russian and Ukrainian governments. To achieve this, we systematically gather tweets posted by accounts affiliated with these governments, the comprehensive list of which is detailed in Table 2. Our collection encompasses 5,993 tweets from Russian government representatives and 9,451 tweets from Ukrainian government representatives, collected over the period between January 24^{th}, 2022, to April 4^{th}. This dataset proves instrumental in establishing our measure of media slant, offering a direct insight into the narratives each government endorsed and ensuring our analytical framework's robustness.

Figure 2 provides insights into the content of these government tweets through keyword frequency analysis. We differentiate the frequencies of Russian and Ukrainian government tweets, represented in

Figure 1: *Geographical distribution of users*



Notes: The figure shows the geographical distribution of users in our sample. In order to locate users in the countries of our analysis, Austria, France, Germany, Italy, Ireland, Switzerland and United Kingdom, we follow the geo-location pipeline proposed in Gehring and Grigoletto (2023).

purple and orange, respectively. Keywords like 'aggression' and 'invasion' are predominantly used by Ukrainian accounts to portray the conflict as an invasion, contrasting with the Russian portrayal as a 'military operation'. Other stems like 'occupi', 'defense', 'nato', 'west', 'nazi', and 'donbass' further delineate the narratives of each side. The use of these terms underlines the slant in the content from these government accounts, making them suitable benchmarks for our measurement.

Following Gennaro and Ash (2023) we take all Ukrainian tweets in GT, create a vector representation using the text embedding model sentence-t5-xl (Ni et al. 2021), and average those representations to produce a single vector representing the Ukrainian government pole. We compute the Russian government pole analogously. Then, we embed all UT of our main analysis' dataset (see data section 2) with sentence-t5-xl and use Equation 1 to obtain a score for each input tweet. This score is a ratio measuring the language similarity between the given tweet and the Russian pole relative to the similarity between the given tweet and the Ukrainian pole. Formally, we compute this ratio as follows:

$$Y = \frac{sim(d, R) + b}{sim(d, U) + b} - 1, \tag{1}$$

where d denotes a vector representing the input text, R and U are vectors representing the two poles, b is a smoothing parameter set to 1, and sim refers to the cosine similarity. We subtract 1 to center the ratio

Table 2: *Accounts of the Russian and Ukrainian governments' representatives*

Ukrainian Accounts	Account Holder	Russian Accounts	Account Holder
https://twitter.com/DI_Ukraine	Defence Intelligence	https://twitter.com/RussianEmbassy	Embassy in the UK
https://twitter.com/Ukraine	Ukraine	https://twitter.com/mfa_russia	Ministry of Foreign Affairs
https://twitter.com/DefenceU	Ministry of Defense	https://twitter.com/mission_rf	Mission to the International Organizations in Vienna
https://twitter.com/CinC_AFU	Colonel General Oleksandr Syrskyi	https://twitter.com/RF_OSCE	Mission to the OSCE
https://twitter.com/oleksiireznikov	Minister of Defence	https://twitter.com/RusEmbUSA	Embassy in the US
https://twitter.com/kabmin_ua_e	Cabinet of Ministers	https://twitter.com/RussianEmbassyC	Embassy in Canada
https://twitter.com/MFA_Ukraine	Ministry of Foreign Affairs	https://twitter.com/KremlinRussia_E	Official Kremlin News
https://twitter.com/DmytroKuleba	Minister of Foreign Affairs	https://twitter.com/EmbassyofRussia	Embassy in South Africa
https://twitter.com/AndriyYermak	Head of the Office of the President	https://twitter.com/PMSimferopol	Ministry of Foreign Affairs' Office in Crimea
https://twitter.com/NSDC_ua	Press Service of the National Security and Defense Council	https://twitter.com/RusMission_EU	Mission to the EU
https://twitter.com/UKRinDEU	Embassy of Ukraine in Germany	https://twitter.com/RusBotschaft	Embassy in Germany
https://twitter.com/ukrinche	Embassy of Ukraine in Switzerland	https://twitter.com/RusEmbSwiss	Embassy in Switzerland
https://twitter.com/ukrinfra	Embassy of Ukraine in France	https://twitter.com/ambrusfrance	Embassy in France
https://twitter.com/ukrinit	Embassy of Ukraine in Italy	https://twitter.com/rusembitaly	Embassy in Italy
https://twitter.com/UkrEmbLondon	Embassy of Ukraine in the UK		
https://twitter.com/MelnykAndrij	Ukrainian Ambassador to Germany		

Notes: The table reports the Russian and Ukrainian government-affiliated accounts that were used as sources for the two poles used to create our slant measurement. For each of these accounts, we extracted tweets in English between January 24^{th}, 2022, and April 4^{th}, 2022. This extraction resulted in 5,993 tweets for the Russian pole and 9,451 tweets for the Ukrainian pole. Note that to increase the number of English accounts on the Russian side, we also included the embassy account for non-European English-speaking countries active on Twitter.

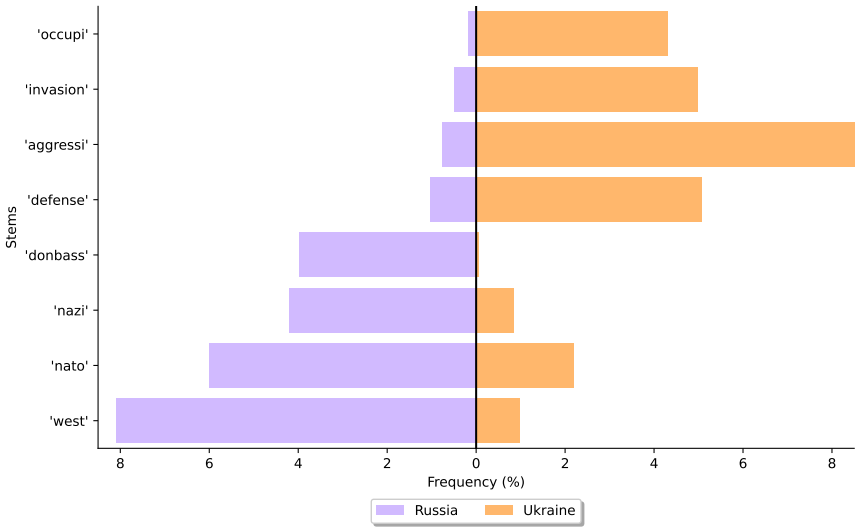
around zero. Positive values indicate tweets more similar to the Russian pole. Tweets with negative values are more similar to the Ukrainian pole. If $Y = 0$, this means that the input text is equally similar to the Ukrainian and Russian poles. We compute the poles as time-varying measures to account for possible changes over time in the official viewpoints. The comparison poles for a day t consist of all government tweets between days t-7 and t, using a decay factor of 0.5 to reduce the influence of more distant days[8]. The pole ratio for each user tweet from day t is then computed based on the two corresponding daily poles.

Finally, we standardize the resulting media slant ratio to a mean of 0 and a standard deviation of 1. Increasing our final measure by one unit implies moving one standard deviation closer to the Russian pole. Appendix Figure B.1 shows examples of tweets and their corresponding score. When in the following sections we refer to tweets and retweets of pro-Russian slant, we refer to a binary measure. This measure is obtained by assigning 1 to any tweet or retweet whose ratio score is one standard deviation above 0, hence towards the Russian pole, and 0 otherwise. We use the threshold of one standard deviation to mitigate the inclusion of potential noise around the mean value of 0. Nevertheless, we perform robustness checks using a zero threshold to further ensure the reliability of our findings.

In Figure 3, we present the daily average measure of media slant in both EU and non-EU countries as part of our analysis. The graph reveals several notable trends. In the days marking the onset of the invasion, both EU and non-EU regions reached a minimum in average slant, suggesting a widespread initial reaction leaning towards the Ukrainian pole. Subsequently, there is a pronounced and consistent shift towards the Russian pole, underscoring the European Commission's concern that the conflict was being waged not only on the ground but also online, with the EU particularly targeted by Russian propaganda

[8] The decay factor of 0.5 results in the following weights: 0.5, 0.55204476, 0.60950683, 0.6729501, 0.74299714, 0.82033536, 0.90572366, 1.

Figure 2: *Word frequency in the sample of government tweets*



Notes: The figure shows frequencies for selected word stems in the sample of government tweets. In purple, we show the frequency in tweets coming from representatives of the Russian government, and in orange for the Ukrainian government. Frequencies represent the percentage of tweets containing the stem of each specific word of interest. Results are based on 9,451 tweets from Ukrainian government exponents and 5,993 tweets from Russian government exponents.

efforts. Furthermore, the movement of average slant in EU and non-EU countries shows similar trends before the ban's implementation, after which a distinct divergence is observed. This pattern may reflect the ban's impact, a hypothesis we will explore more systematically in the subsequent sections. In Appendix Figure B.2 we show the same using only original tweets.

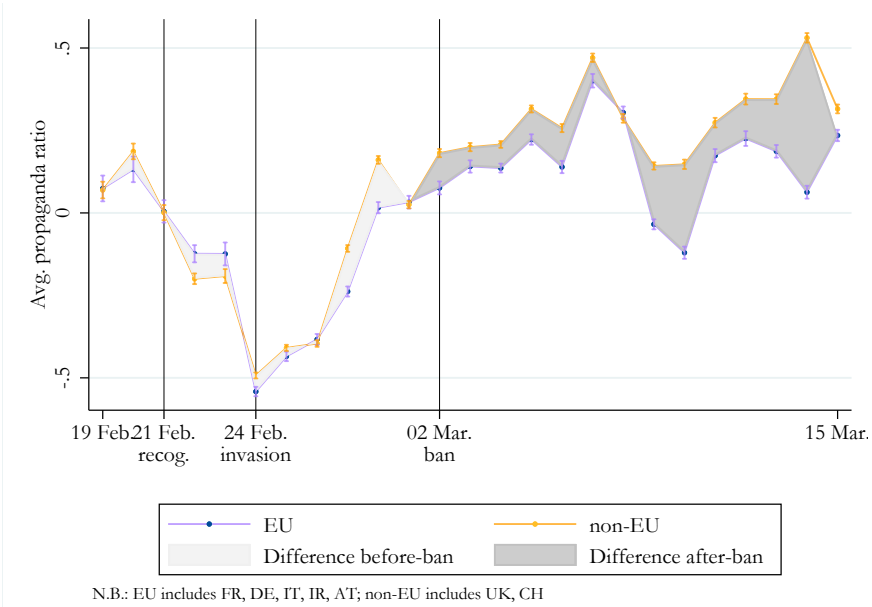
4 Conceptual framework and empirical method

4.1 A supply shock to the market for news

The enforcement of the ban against Russia Today and Sputnik represents a significant shock to the news market. With millions of followers and a presence across diverse online platforms, they were central nodes in spreading state-aligned Russian narratives. Their direct links to the Russian government enhanced their effectiveness, enabling them to access and disseminate information rapidly and widely. Therefore, their sudden block within the EU from a crucial platform like Twitter did not just eliminate two prolific voices, it fundamentally disrupted the network of information flow. This disruption is akin to removing a major supplier from a traditional market, with expected repercussions on both the availability of certain narratives and the overall dynamics of information dissemination.

We follow the literature on media bias (Mullainathan and Shleifer 2005; Gentzkow and Shapiro 2010) and think of the news market about a specific subject, the Russo-Ukrainian conflict, as a Hotelling-type market structure. Consumers are distributed along a line of beliefs and have a preference to read news consistent with their beliefs. The two ends of the distribution represent the most extreme worldviews, in our case, the Russian and Ukrainian governments. News outlets cater to a particular segment of

Figure 3: *Time-series of our slant measure: Daily averages*



Notes: The figure shows the daily averages of our media slant measurement, in the time-frame of our analysis, between February 19^{th}, 2022 to March 15^{th}, 2022. The measure is normalized to have a mean of 0 and a standard deviation of 1. When positive the measure indicates content closer to the Russian pole, and when negative it indicates content closer to the Ukrainian pole. In purple, we show the daily averages in the EU countries in our study, Austria, France, Germany, Ireland, and Italy, while in orange the daily averages in non-EU countries, United Kingdom and Switzerland. In grey, the difference between the two averages. We indicate the following important dates: 21^{st} Feb. for the official recognition by Putin of the Donetsk People's Republic and the Luhansk People's Republic, 24^{th} Feb. as the beginning of the war and 2^{nd} Mar. as the beginning of the ban. Appendix Figure B.2 reproduces the same time-series but limiting the analysis to original tweets excluding retweets.

consumers along this distribution. The choice of the segment is driven by two motives: medium-internal preferences over worldviews and maximizing outreach.

In this setup, the ban takes the form of a classic supply shock. Users who previously read and interacted with Russia Today and Sputnik lose a source of information[9]. In a traditional news market, the affected consumers would resort to other news outlets. Additionally, outlets could adjust their reporting in an attempt to capture this vacated share of readers. However, on social media, we expect users not only to adjust their news consumption behavior but also their production of content by posting and sharing information. Notably, in the very short run, the ban leads to a mechanical decrease in the amount of pro-Russian slant content that can be shared. However, social media also allows users not affiliated with any media outlet to share or create information aligning with their worldview. Therefore, certain users may try to fill the information gap created by the ban. We test this hypothesis in Section 5.4.

To assess the ban's impact, we examine the dynamics of pro-Russian and pro-Ukrainian slant content on Twitter from three analytical angles. The first aspect is the intensive margin or the depth of intensity of the slant content. The ban might signal a broader intolerance towards extreme narratives, potentially prompting users who engage with or disseminate such content to re-calibrate their approach. This could

[9] Importantly, the way Twitter enforced the ban did not allow for direct circumvention of the ban via a VPN on an account created before the ban. We address the potential for circumvention by creating new accounts after the ban in Appendix F.

lead to a decrease in the overall extremism of tweets, as users might moderate their language and content to avoid being flagged or banned.

The second aspect, the extensive margin, concerns the volume of slanted content production, which we quantify by counting the number of original pro-Russian slant tweets with our binary measure. With focal *suppliers* of Russian-aligned narratives removed, users who previously directly or indirectly relied on these sources might experience an increased cost of searching for pro-Russian content, resulting in reduced pro-Russian content creation. In turn, this could also spur users to compensate for the loss of ready-made content by increasing their production of original tweets that align with or support the now-absent narratives. Such a response would reflect an attempt to maintain the presence of these narratives in the public discourse, despite the absence of their primary propagators. The empirical question we tackle in our analysis is which of the two forces dominates.

The third aspect focuses on the spread of slant content, particularly through retweets, and can still be conceived as part of the extensive margin. By cutting off a significant content source, the ban will likely impact the volume of retweets of pro-Russian content. Users finding fewer original posts from the banned outlets directly or from other users building on banned outlet content to share might reduce their retweeting behavior. This could result in a noticeable decline in the spread of Russian-aligned narratives on the platform, affecting the reach and penetration of these narratives among the wider audience.

By examining these three aspects – intensity, production, and spread – we aim to comprehensively understand the multifaceted effects of the ban. Each dimension provides a different lens through which to view the ban's impact, collectively offering a holistic picture of how a significant policy reshapes the digital landscape of media slant and information dissemination.

4.2 Identification strategy

We use a difference-in-difference design to estimate the causal effect of banning the Russian outlets by comparing tweets posted by users in the EU, who were affected by the ban, to tweets posted by users outside the EU, whose exposure to these outlets was not restricted. We estimate difference-in-difference specifications at the user-time level of the following form:

$$Y_{i,t} = \alpha_i + \gamma_t + \beta EU_i \times Ban_t + \Theta \boldsymbol{X}_{i,t} + \epsilon_{i,t}, \tag{2}$$

where $Y_{i,t}$ is a measure of user behavior, such as the average media slant ratio measure, the number of pro-Russian slant tweets, and retweets, by user i in period t. These outcomes target the arguably most policy-relevant issues at the core of the EU's justification of the ban: the overall qualitative content of the discourse on the war in Ukraine and the quantity of pro-Russian narratives spread in the online information space. EU_i is an indicator variable equal to 1 for users located in the European Union and 0 otherwise. Ban_t is an indicator equal to 1 after March 2^{nd}, 2022. We also estimate an event study version of Equation 2 to investigate the timing of the effect. α_i and γ_t are a full set of user and day-fixed effects absorbing average differences in tweet content across users and time. $X_{i,t}$ is a vector of additional

control variables capturing tweet style information, i.e., the number of words, hashtags, and mentions.

To interpret β in Equation 2 as the causal effect of the ban on Russia Today and Sputnik, we require the assumption that Twitter user behavior in EU and non-EU countries would have followed parallel trends in the absence of the ban. While this assumption is not directly testable, we provide two pieces of evidence to support it. First, we test for balance in tweet characteristics *before* the ban between users located in the EU and non-EU countries in our sample. Appendix Figure B.5 shows the results of this exercise. We detect some differences and we control for the variables mentioned above. These variables do not affect our estimates. Second, in an event study specification, we investigate the link between user behavior and being located in the EU. This allows us to see whether the slant of users in the EU compared to users outside the EU followed similar trends before the ban. The results of this exercise show no meaningful differential trends in outcomes before the ban, making it less likely that the trends would have diverged in the absence of the ban.

Another potential concern is that the invasion itself might impact the use of pro-Russian or pro-Ukrainian slant on Twitter. Day-fixed effects absorb any common shock to EU and non-EU countries. We only include major Western European countries in our sample to make it less likely that differential exposure to the war itself or fear of spillover to the country of residence confounds our estimates. Eastern European countries bordering Russia or Ukraine might have a differential response to the war, so we excluded them from the analysis.

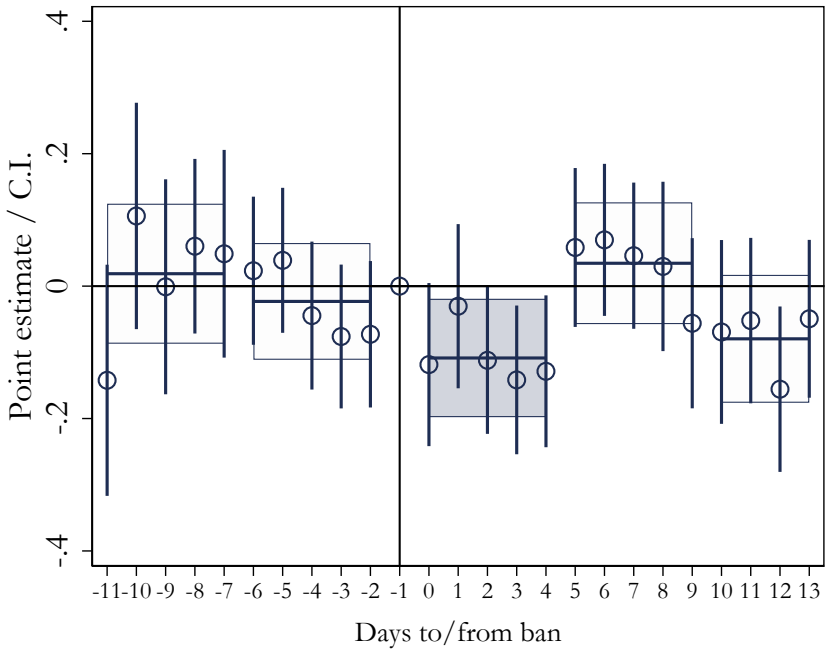
A final concern is that the ban was not enforced and a clean assignment of the treatment status in our setting also depends on the absence of ways to circumvent the ban. As outlined in detail in Appendix B, we use the location of Twitter users as indicated in their profiles to assign them to a country and thereby a treatment status. These locations are public information and can be retrieved via the Twitter API. However, Twitter also gathers non-public information on users' locations to determine the content that cannot be displayed to a user. According to Twitter's public documentation, such information does not only include IP addresses, which can be easily changed via a VPN but, for example, also wireless networks or cell towers near a user. Crucially, manually changing this non-public country setting does not change the content withheld by Twitter due to local laws. Therefore, even if readers of RT and Sputnik who reside in one of the countries that enacted the ban and had their location assigned by Twitter use a VPN to reach the website of the outlets, they still cannot access, read, or interact with any account of RT and Sputnik.

5 Results: The impact of the ban

5.1 Direct effect

The first step of our analysis involves assessing the impact of the ban on individuals who previously engaged with Russia Today and Sputnik. These users – that from now on we indicate as the *interaction users* – are pivotal for several reasons. First, the ban directly impacts them by removing a critical source of information and news content from their online social network. Second, while they represent a minor segment of our sample, their contribution to the overall volume of tweets is substantial. Third, this group

Figure 4: *Daily event-study on our slant measure: Interaction users*



Notes: The figure displays coefficients and 95% confidence intervals from estimating a daily event study version of Equation 2. The sample consists of *interaction users* – users who interacted with the banned outlets before the ban – and includes users located in the UK and Switzerland as control group and users located in Austria, France, Germany, Ireland, and Italy as treatment group. The dependent variable is each user's daily average of slant in tweets. We use tweets from the period between 19^{th} February to 15^{th} March 2022. We estimate Equation 2 including user- and day-fixed effects relative to the omitted day, 1^{st} March 2022 immediately before the ban's implementation, and controlling for word count, mentions count, and hashtags count. In the aggregate specification, coefficients of interest are interactions between a dummy variable for aggregated intervals for 19^{th} to 23^{th} February, 24^{th} to 28 Februaryth, 2^{nd} to 6^{th} March, 7^{th} to 11^{th} March and 12^{th} to 15^{th} March, relative to the omitted day, 1^{st} March 2022 immediately before the implementation of the ban. Coefficient estimates on the day interactions are plotted as dots with their 95% confidence intervals indicated with vertical lines. Coefficient estimates on the aggregate interactions are shown with horizontal lines, and their 95% confidence intervals are indicated as boxes. We cluster standard errors at the user level. Appendix Figure C.1 displays regression results using alternative estimators proposed by Callaway and Sant'Anna (2021) and Borusyak, Jaravel, and Spiess (2024). Appendix Figures D.1 and D.2 reproduce the same results using four different alternative models of vectorization. Appendix Figures E.1, F.1 show the same analysis excluding users that are plausible bots and excluding accounts created only after the ban, respectively.

predominantly disseminated pro-Russia content prior to the ban – as shown in Appendix Figure B.3 – placing them at the core of the EU's regulatory measure.

Figure 4 shows results of estimating a daily event study version of Equation 2 with our media slant measure as the dependent variable. This approach captures what we call the intensive margin of the effect, focusing on shifts in content slant. The analysis solely includes tweets from February 19^{th}, 2022, to March 15^{th}, 2022, produced by the *interaction users*. Each regression incorporates user- and day-fixed effects, with standard errors clustered at the user level. In our main specification, we estimate regressions using OLS. Appendix C provides results using alternative estimators proposed by Callaway and Sant'Anna (2021) and by Borusyak, Jaravel, and Spiess (2024) and outlines some considerations on the use of the different estimators in our setting.

Figure 4 provides three key insights. First, we do not observe discernible pre-trends in the period

leading up to the ban. Second, immediately after the ban's implementation, we observe a negative shift indicating a movement toward the Ukrainian pole or, at the very least, a move away from the most extreme pro-Russian positions. This shift is statistically significant, with a point estimate of 0.11 standard deviations for the aggregated estimate of the first five days after the ban. Third, the effect is notably very short-lived, with differences between treated and untreated users reverting to being closer to zero and not statistically significant from day six onwards. Alternative DID estimators presented in Figure C.1 show a similar pattern and suggest somewhat stronger effects.

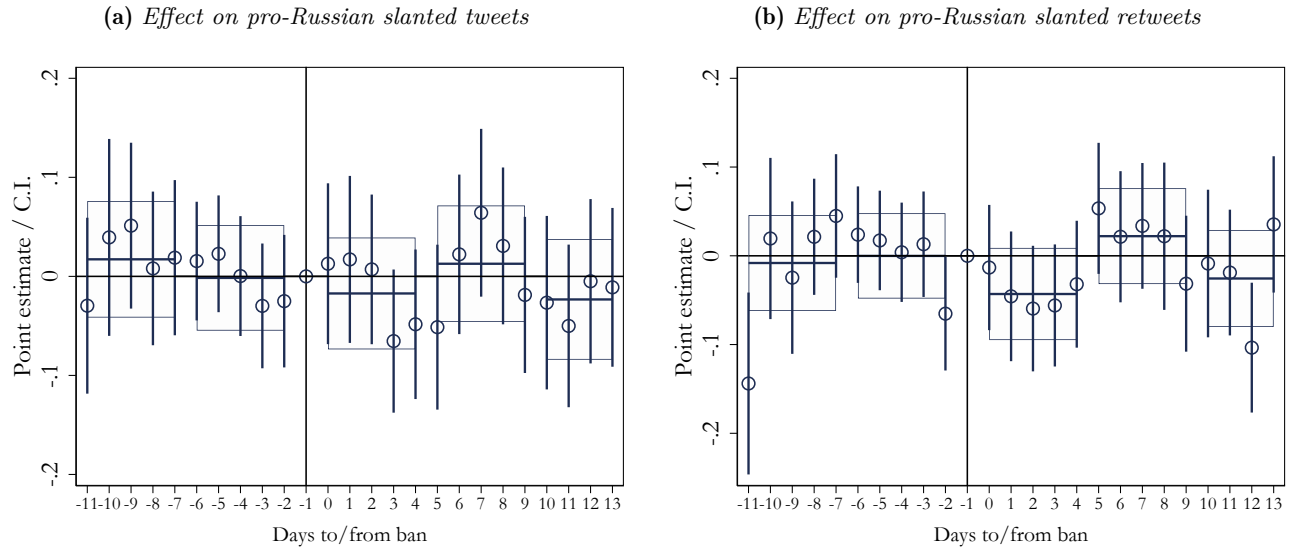
After examining the impact on the content's slant, our analysis next focuses on the extensive margin. We investigate whether there is a change in the total volume of pro-Russia content produced by *interaction users* in the European Union compared to those located outside the EU. In this part of our study, we define a tweet or retweet as pro-Russia if our media slant measure places it one standard deviation above 0 towards the Russian pole. This threshold is chosen to minimize the risk of including tweets that might be mistakenly identified as pro-Russia due to only slight variations in the slant measure. For robustness, we also perform analyses with a threshold of 0, as detailed in Appendix G.

Figures 5a and 5b display the results of exploring the effects of the ban on the extensive margin. Specifically, Figure 5a evaluates the ban's impact on the share of a user's total tweets classified as pro-Russia. Similarly, Figure 5b assesses how the ban influenced the share of retweets by a user that falls into the pro-Russia category. In contrast to the insights derived from the intensive margin analysis, we find that shares of tweets and retweets identified as pro-Russia do not indicate a noticeable effect of the ban. Similarly to the previous results, no clear pre-trend is observable in the period preceding the ban.

Next, we delve deeper into the temporary nature of the ban's impact on the *interaction users*. Table 3 displays results of regressions using specification 2, with additional interaction terms for each of the two weeks after the ban. Column 1 shows the effect of the ban in the two weeks after the ban itself on the average slant measure, the same dependent variable used for results in Figure 4. Columns 3 and 4 show respectively the effect on the share of pro-Russia tweets and pro-Russia retweets out of all tweets produced in a day by a user (and captured by our query), respectively the same dependent variables considered in Figure 5a and 5b. Columns 4 and 5 show the impact of the total number – instead of share – of pro-Russia tweets and retweets produced daily by each user.

The findings offer a consistent picture across all measures of the prevalence of Russian propaganda amongst *interaction users*. We find a statistically significant reduction in the average slant and volume of Russian propaganda created and spread in the first week after the ban's implementation. As shown in Column 1, the media slant decreased by 73.8% from its pre-ban average, a statistically significant change. Columns 2 and 3 illustrate that the point estimates suggest that the share of pro-Russia tweets and retweets declined by approximately 17.4% and 10.2%, respectively, compared to their pre-ban averages – both estimates are, however, not statistically significant. While more modest, the reduction in the absolute number of pro-Russia tweets by 3.1%, documented in Column 4, is statistically significant in the first week after the ban. Column 5 shows that our point estimates suggest a statistically not significant 1.8% reduction in the absolute number of pro-Russia retweets. However, all effects diminish and lose statistical significance in the second week, suggesting a short-lived effect of the ban.

Figure 5: *Daily event-study on share of slanted tweets and retweets: Interaction users*



Notes: The figure displays coefficients and 95% confidence intervals from estimating a daily event study version of Equation 2. The sample consists of *interaction users* – users who interacted with the banned outlets before the ban – and includes users located in the UK and Switzerland as control group and users located in Austria, France, Germany, Ireland, and Italy as treatment group. We use tweets from the period between 19^{th} February to 15^{th} March 2022. We estimate Equation 2 including user- and day-fixed effects relative to the omitted day, 1^{st} March 2022 immediately before the ban's implementation, and controlling for word count, mentions count, and hashtags count. In the aggregate specification, coefficients of interest are interactions between a dummy variable for aggregated intervals for 19^{th} to 23^{th} February, 24^{th} to 28 Februaryth, 2^{nd} to 6^{th} March, 7^{th} to 11^{th} March and 12^{th} to 15^{th} March, relative to the omitted day, 1^{st} March 2022 immediately before the implementation of the ban. Coefficient estimates on the day interactions are plotted as dots with their 95% confidence intervals indicated with vertical lines. Coefficient estimates on the aggregate interactions are shown with horizontal lines, and their 95% confidence intervals are indicated as boxes. We cluster standard errors at the user level. Figure 5a display estimates using user's daily share of pro-Russian slanted tweets – defined as having a media slant measure above 1 – as dependent variable. Figure 5b display estimates using user's daily share of pro-Russian slanted retweets – defined as having a media slant measure above 1 – as dependent variable. Appendix Figure C.2 displays regression results using alternative estimators proposed by Callaway and Sant'Anna (2021) and Borusyak, Jaravel, and Spiess (2024). Appendix Figures E.2, F.2 and G.1 show the same analysis excluding users that are plausible bots, excluding accounts created only after the ban and using 0, instead of 1, threshold to define the binary variables of pro-Russia tweets and retweets, respectively.

In sum, we document a statistically significant change in the average slant of content within the EU shifting away from the Russian pole, indicating an effect at the intensive margin in the time frame of our sample. This overall reduction is primarily driven by an immediate response in the first week after the ban and is reduced quickly in the second week after the ban. At the extensive margin, we find only small effects on the volume of clearly pro-Russian content that are not statistically significant and, if anything, strongest in the first week after the ban. This discrepancy suggests that the ban's overall impact on users who interacted with the banned outlets is subtle, with a short lived effect on the average slant of content but no meaningful and statistically significant impact on the overall level of clearly pro-Russian propaganda generated and spread by *interaction users*.

In the next step, we explore the heterogeneity of the ban's effects on users with different levels of slant in their pro-Russian content before the ban to provide a clearer understanding of how content adjustments span across the spectrum of user types.

Table 3: *User-day level two-periods TWFE with post-ban weeks interactions: Interaction users*

	(1) Avg. media slant	(2) % pro-Russia tweets	(3) % pro-Russia retweets	(4) Tot. Pro-Russia tweets	(5) Tot. pro-Russia retweets
EU × 1st week after-ban	-0.050	-0.020	-0.017	-0.041	-0.034
	[0.023]	[0.014]	[0.013]	[0.020]	[0.026]
EU × 2nd week after-ban	-0.034	-0.002	-0.010	0.004	-0.018
	[0.025]	[0.016]	[0.014]	[0.017]	[0.024]
User FEs	yes	yes	yes	yes	yes
Day FEs	yes	yes	yes	yes	yes
Observations	29704	16508	19614	29704	29704
R^2	0.343	0.236	0.247	0.215	0.375
Pre-period mean of DV	-0.068	0.113	0.162	1.324	1.861
1st week % of mean	-73.83	-17.39	-10.23	-3.07	-1.83

Notes: The table displays coefficients from estimating Equation 2 examining the ban's differential impact in the two weeks following the ban, on users who interacted with Russia Today and Sputnik before the ban. Column 1 shows the effects on our measure of media slant, the intensive margin. Columns 2 and 3 show effects on our measure of the extensive margin, respectively, the share of pro-Russia tweets and retweets. Columns 4 and 5 show the effect on the total number of pro-Russia tweets and retweets produced by the author in the time period. The sample consists of *interaction users* – users who interacted with the banned outlets before the ban – and includes users located in the UK and Switzerland as control group and users located in Austria, France, Germany, Ireland, and Italy as treatment group. We use tweets from the period between 19^{th} February to 15^{th} March 2022. We estimate Equation 2 including user- and day-fixed effects, and controlling for word count, mentions count, and hashtags count. Standard errors clustered at the user level are reported in brackets. Appendix Tables E.2, F.2 and G.2 show the same analysis excluding users that are plausible bots, excluding accounts created only after the ban and using 0, instead of 1, the threshold to define the binary variables of pro-Russia tweets and retweets, respectively.

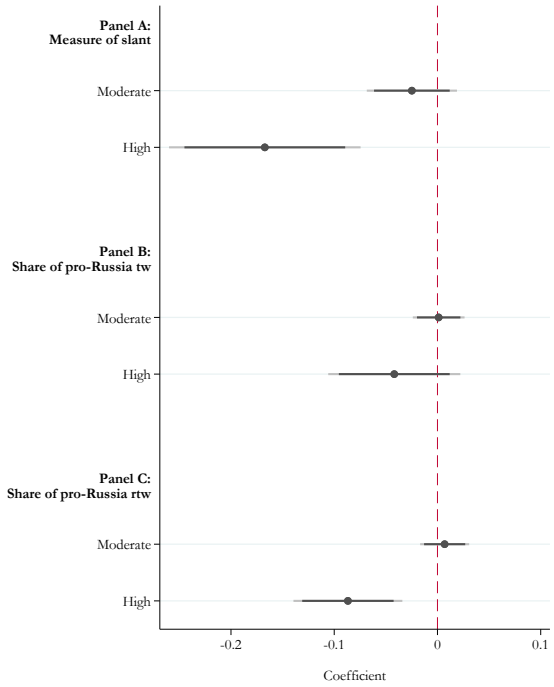
5.2 Heterogeneous effect among *interaction users*

Figure 6 illustrates the ban's heterogeneous effects, based on the average level of the pro-Russian slant of users interacting with the outlets before the ban. We categorize *interaction users* into two subgroups for analysis using the specification outlined in Equation 2: a high pro-Russian slant group, identified as those in the top 25^{th} percentile of the slant distribution, and a moderate group, encompassing the remaining 75% of the distribution. Panel A addresses the intensive margin, reflecting changes in the average media slant. Panels B and C examine the extensive margin, focusing on the volume of pro-Russia tweets and retweets produced by these groups. The figure shows the coefficients with 90% and 95% confidence intervals, the latter in light grey, using standard errors clustered at the user level.

We find that users with a moderate level of pro-Russia slant before the ban display no significant shifts across the measured dimensions. Importantly, these are always relative changes to the corresponding control group – moderate users who used to interact with the outlets – in the non-EU countries in our analysis. In contrast, users who, before the ban, displayed the highest levels of average pro-Russia slant demonstrate a marked response to the ban. In particular, these users show a decrease in both the intensive margin captured by the average media slant post-ban and the extensive margin captured by the number of pro-Russia retweets post-ban. Neither moderate group users nor the most extreme ones display a statistically significant response in their production of own pro-Russia tweets.

Taken together, these results seem to indicate that the ban had an effect on those *interaction users* that used to engage with the outlets – Russia Today and Sputnik – before the ban. Despite being short-lived, the point estimates suggest a meaningful effect. In particular, the ban mainly had an impact on the

Figure 6: *Heterogeneous effects of the ban by pre-ban level of pro-Russian slant: Interaction users*



Notes: The figure presents coefficients from estimating Equation 2 assessing the ban's heterogeneous impact on users that used to interact with the outlets before the ban. Users are divided into two groups: moderate when the average slant in their tweets before the ban is in the bottom 75% of the distribution, and high when the average slant in their tweets before the ban is in the top 25%. The coefficients are shown with 90% and 95% confidence intervals (95% in light grey). Panel A shows the results of our measure of media slant, the intensive margin, obtained by taking daily averages of media slant for each user. Panel B and C show results on the daily proportion of tweets/retweets classified as pro-Russia, out of all tweets/retweets produced by the user and captured by our query. The sample consists of *interaction users* – users who interacted with the banned outlets before the ban – and includes users located in the UK and Switzerland as control group and users located in Austria, France, Germany, Ireland, and Italy as treatment group. We use tweets from the period between 19^{th} February to 15^{th} March 2022. We estimate Equation 2 including user- and day-fixed effects, clustering standard errors at the user level, and controlling for word count, mentions count, and hashtags count. Appendix Figures E.3, F.3 and G.2 show the same analysis excluding users that are plausible bots, excluding accounts created only after the ban and using 0, instead of 1, threshold to define the binary variables of pro-Russia tweets and retweets, respectively.

intensive margin – shifting the slant of the content rather than the amount of it. The shift seems to have mainly occurred among those *interaction users* that, before the ban, were producing and spreading highly slanted pro-Russian content. There are several potential explanations for the observed pattern. These users may have been more actively engaged with the content from the banned outlets or heavily relied on these sources for their information. This interpretation is supported by the fact that these highly pro-Russian users also experienced a reduction in the extensive margin, notably in the percentage of pro-Russia retweets, likely due to the removal of two significant sources from their information network. An alternative explanation could be that the ban acted as a signal, making potential repercussions of disseminating extremely pro-Russian content more tangible. This interpretation would suggest that the most extreme users were the ones most influenced by the signalling effect of the ban.

5.3 Indirect effect

In the previous sections, we explored how the ban affects media slant and the volume of pro-Russia content among the *interaction users*. To give deeper insight into the reach of the ban beyond the users directly affected, we explore the indirect effect on users who had not interacted directly with the banned outlets – we label these as *non-interaction users*. It is crucial to clarify that this absence of interaction does not imply that these users were disengaged from creating, consuming, or sharing pro-Russia content. Instead, their involvement with such content did not occur through direct engagement with the outlets now under the ban. The ban may indirectly influence them through the spread of the outlet's content in the social media network. This distinction allows us to explore the ban's broader effects on the online discourse beyond the immediate circle of the banned outlets' direct users.

Table 4 reports results of regressions of estimating Equation 2. We report results for the *interaction users* – in Panel A – and for *non-interaction users* – in Panel B. In both panels, Column 1 shows the effect of the ban on our measure of content slant. Columns 3 and 4 show, respectively, the effect on the share of pro-Russia tweets and pro-Russia retweets out of all tweets produced in a day by a user. Columns 4 and 5 show the impact of the total number – instead of share – of pro-Russia tweets and retweets produced daily by each user.

As already shown above, the ban had a strong impact in decreasing media slant among users who used to interact with the outlets before the ban. However, we also find an effect, albeit less pronounced, for users only indirectly affected by the ban. Compared to the 63.1% reduction relative to the pre-ban average, the *non-interaction users* only display a decrease of 17.3% in media slant. It is important to notice that this is still a statistically significant effect despite being more moderate. In turn, there is no effect on either group of users for both the share and the absolute number of pro-Russia tweets. Interestingly, pro-Russia retweets display a more pronounced decrease in relative terms among the users who did not interact with the outlets before the ban. In fact, the decrease among these users is about 26.9% relative to the pre-ban average and statistically significant on a 5% level, while it is 8.5% the pre-ban average for *interaction users* and not statistically significant.

Some key observations stand out when assessing the ban's overall effectiveness. First, while the ban initially had a pronounced effect on users who were previously engaging with the outlets, this impact was notably short-lived. Second, the ban's indirect effect on *non-interacting users* was comparatively milder, especially regarding the intensity of the media slant. Third, it is crucial to interpret these findings within the limited time-frame of our study. Our data stops on March 15^{th}, 2022, just as the UK joined the European ban. Despite the brevity of this period, there was an immediate adjustment in the media market following the ban. This rapid adaptation suggests a dynamic response to regulatory measures, although the long-term implications are beyond the scope of this paper.

5.4 Mechanisms

Moving forward, our analysis aims to delve into potential mechanisms within the Twitter media market that might explain the observed limited and short-lived impact of the ban. Specifically, in this section, we explore the role of *suppliers* of slanted content. We define a *supplier* of slanted content as any user

Table 4: *User-day level two-periods TWFE: Interaction and non-interaction users*

Panel A: Interaction users

	(1) Avg. media slant	(2) % pro-Russia tweets	(3) % pro-Russia retweets	(4) Tot. Pro-Russia tweets	(5) Tot. pro-Russia retweets
EU × after-ban	-0.043	-0.012	-0.014	-0.021	-0.027
	[0.020]	[0.012]	[0.011]	[0.015]	[0.021]
User FEs	yes	yes	yes	yes	yes
Day FEs	yes	yes	yes	yes	yes
Observations	29704	16508	19614	29704	29704
R^2	0.343	0.236	0.247	0.215	0.375
Pre-period mean of DV	-0.068	0.113	0.162	1.324	1.861
% of mean	-63.13	-10.88	-8.45	-1.60	-1.45

Panel A: Non-interaction users

	(1) Avg. media slant	(2) % pro-Russia tweets	(3) % pro-Russia retweets	(4) Tot. Pro-Russia tweets	(5) Tot. pro-Russia retweets
EU × after-ban	-0.034	0.002	-0.038	-0.004	-0.011
	[0.007]	[0.004]	[0.004]	[0.006]	[0.005]
User FEs	yes	yes	yes	yes	yes
Day FEs	yes	yes	yes	yes	yes
Observations	312779	147536	181353	312779	312779
R^2	0.424	0.328	0.313	0.299	0.297
Pre-period mean of DV	-0.199	0.101	0.140	0.934	1.110
% of mean	-17.27	1.75	-26.85	-0.44	-1.00

Notes: The table displays coefficients from estimating Equation 2 examining the ban's impact on users who interacted with Russia Today and Sputnik before the ban in Panel A, and on user that had no interactions with the outlets in Panel B. Column 1 shows the effects on our measure of media slant, the intensive margin. Columns 2 and 3 show effects on our measure of the extensive margin, respectively, the share of pro-Russia tweets and retweets. Columns 4 and 5 show the effect on the total number of pro-Russia tweets and retweets, respectively, produced by the author in the time period. The sample includes users located in the UK and Switzerland as control group and users located in Austria, France, Germany, Ireland, and Italy as treatment group. We use tweets from the period between 19^{th} February to 15^{th} March 2022. We estimate Equation 2 including user- and day-fixed effects, and controlling for word count, mentions count, and hashtags count. Standard errors clustered at the user level are reported in brackets. Appendix Tables E.3, F.3 and G.3 show the same analysis excluding users that are plausible bots, excluding accounts created only after the ban and using 0, instead of 1, threshold to define the binary variables of pro-Russia tweets and retweets, respectively.

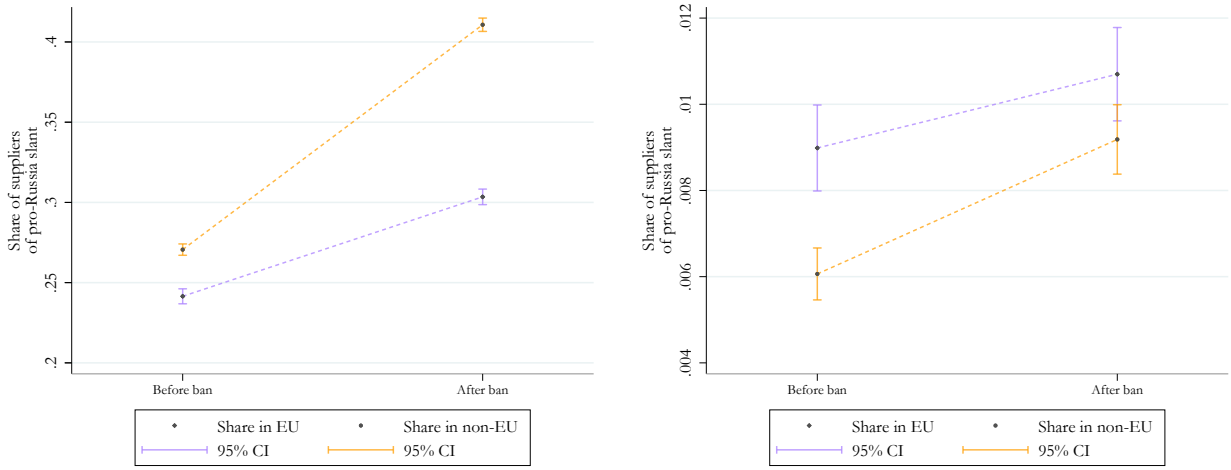
that produced or shared at least one tweet or retweet whose content scores above 1 in our slant measure in the period before the ban. Focusing on these users, we aim to provide suggestive evidence to uncover dynamics that may have counteracted the effect of the ban.

The first mechanism that may counteract the effect of the ban is the emergence of new content *suppliers* in the EU following the ban. This aspect is crucial for understanding the broader implications of the ban, which acts as a supply shock within Twitter's media ecosystem. By removing two major providers of state-backed pro-Russia content, a gap is created in the network. This gap may be filled by new suppliers, who could be motivated by the reduced competition for influence within the realm of slanted content or might aim to counter the ban's objectives by spreading pro-Russian slant themselves. Therefore, we first focus on exploring the entry of new *suppliers* into the market.

Figure 7 shows the share of users identified as *suppliers* in both EU and non-EU regions within our sample before and after the ban's enforcement. Figure 7b narrows this focus to *suppliers* presumed to be bots, based on criteria outlined by Tabassum et al. (2023). In this analysis, a "bot" is characterized by a user profile that ranks in the upper 25% of the distribution of activity levels while being in the lower

Figure 7: *Share of users supplying slanted content*

(a) *Regular users* **(b)** *Potential bots*



Notes: The figure illustrates the proportion of users classified as *suppliers*, defined as those who have posted at least one tweet or retweet with a slant measure exceeding 1, before and after the ban. The shares within our sample are depicted for users in EU countries (in purple) and non-EU countries (in orange). Specifically, Figure 7a presents the proportion of all users meeting this criterion, while Figure 7b focuses on those identified as *suppliers* and considered plausible bots. The data encompasses the time-frame from 19^{th} February to 15^{th} March 2022

25% of the distribution of the "reputation" metric, defined as the ratio of a user's followers to the sum of their followers and followees.

Despite being merely descriptive, there are three main insights from this analysis. First, we observe a notable increase in the share of *suppliers* in both EU and non-EU countries following the ban. This rise likely aligns with the escalation of the conflict, which was intensifying both on the ground and within the digital sphere during its initial stages. Second, despite the share of *suppliers* growing in both areas, the increase is more pronounced in the non-EU countries. Third, only a relatively small share of users are potential bots-suppliers, and despite starting on a higher level in the EU before the ban, the post-ban increase is more substantial in non-EU countries. These observations imply that while the number of *suppliers* within the EU experienced an increase following the ban, this was less substantial compared to that observed in non-EU regions, thus suggesting other mechanisms might be at play.

The second potential mechanism we examine involves an increase in activity among users who were *suppliers* of slanted content prior to the ban. This consideration is based on a simple premise: a subgroup of users exists who, even before the imposition of the ban, were disseminating pro-Russia slanted content. These individuals were likely in competition with the two banned outlets or complementing their efforts. Following the outlets' removal from the network, it is plausible that these users might have tried to fill the resulting void. Our analysis thus focuses on assessing the ban's impact on these users, also exploring potential heterogeneous effects relative to the level of activity they had before the ban.

Focusing exclusively on users who were *suppliers* of slanted content before the implementation of the ban in the EU or non-EU area, we estimate Equation 2. In Figure 8, Panel A illustrates the ban's effect

on our slant measure, Panel B on the share of pro-Russia tweets generated daily by each user, and Panel C on the share of pro-Russia retweets. Within each panel, we present findings for the entire sample of pre-ban *suppliers*, as well as for two distinct subgroups: moderately active suppliers, constituting those in the lower 75% of the activity level distribution pre-ban, and highly active suppliers, represented by those in the upper 25% of the distribution.

We find that the ban, on average, does not significantly alter the share of pro-Russia tweets among European *suppliers* compared to their non-European counterparts. However, it notably affects the average slant of their tweets and their retweet patterns. Analyzing the two subgroups – moderately active and highly active *suppliers* – reveals the nuanced way the ban influenced these users. The observed impact of the ban primarily comes from *suppliers* who exhibited lower activity levels before the ban. In contrast, it seems to have negligible effects on those more actively engaged in disseminating pro-Russia content. This suggests that moderately active *suppliers*, who may have been willing to spread pro-Russia content at a low cost, might have encountered obstacles in their supply network due to the ban, thereby increasing the cost required to find and produce such content. Conversely, the most active *suppliers* before the ban appear to be individuals prepared to sustain their activity levels, even when confronted with higher costs.

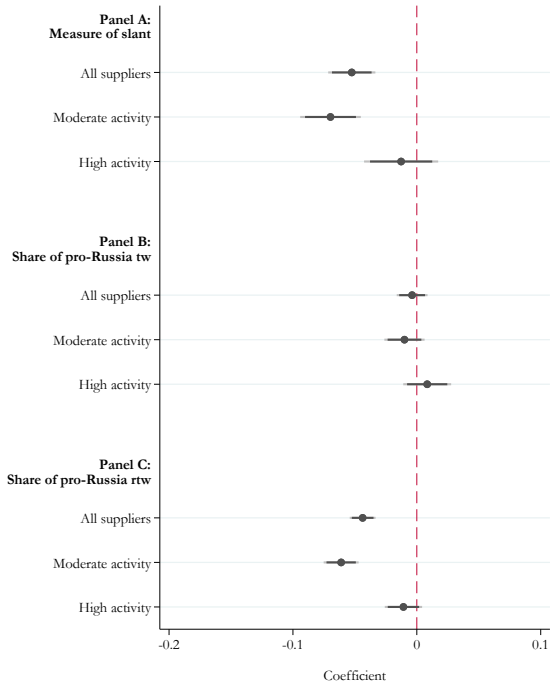
Finally, we ask whether there is a subset of these highly active *suppliers* that not only were not affected by the ban but also reacted to the ban by increasing the levels of activity to counteract it actively. Hence, we focus on the absolute top of this group of users, the 0.5% most active *suppliers* in our sample. This includes only very few suppliers, roughly 50 users, distributed almost equally between EU and non-EU countries. Despite the very low number of users, it is important to mention that these are very active suppliers, with an average activity of 72 posts before the ban, as shown in Appendix Figure B.4.

Table 5 presents results focusing on the top 0.5% of *suppliers* in both EU and non-EU countries, following the estimations used in Table 4. Despite the limitations due to the small sample size, the pattern of response among EU *suppliers* appears to actively counter the ban's intended effects. Across all metrics, the point estimates are consistently positive. Despite the small number of observations, the share of pro-Russia tweets produced by EU *suppliers* compared to their non-EU counterparts post-ban presents a positive and statistically significant coefficient. These findings, while compelling, need to be interpreted with caution due to the limited scope of the dataset. However, it is also critical to acknowledge that our data collection process captures only a subset of overall activity. Therefore, these *suppliers* could exemplify a broader segment actively compensating for the ban's impact, thereby contributing to the resilience of pro-Russia narratives within the Twitter space.

6 Conclusions

Whether and how policymakers should regulate media remains a contentious issue within democratic societies. On one hand, the dissemination of propaganda, fake news, and biased narratives by national and foreign actors poses a substantial threat to the foundation of democratic systems. On the other hand, a core principle of democratic governance is the promotion and protection of freedom of speech. Critics of media regulation are particularly concerned about the potential for censorship and the violation of free

Figure 8: *Heterogeneous effects of the ban by pre-ban activity: Supplier of pro-Russia slant*



Notes: The figure displays coefficients from estimating Equation 2 assessing the ban's heterogeneous impact on users that can be classified as *suppliers*, defined as those who have posted at least one tweet or retweet with a slant exceeding 1 before the ban. We show results for the full sample of pre-ban *suppliers* and for two subgroups of this group. The first sub-group is formed by *suppliers* moderately active before the ban, namely, those whose activity is below the 75^{th} percentile of the pre-ban activity distribution. The second sub-group is formed by *suppliers* highly active before the ban, namely, those whose activity is above the 75^{th} percentile of the pre-ban activity distribution. The coefficients are shown with 90% and 95% confidence intervals (95% in light grey). Panel A shows the results of our measure of media slant, the intensive margin, obtained by taking daily averages of media slant for each user that used to interact with Russia Today and Sputnik before the ban. Panel B and C show results on the daily proportion of tweets/retweets that can be classified as pro-Russia, out of all tweets/retweets produced by the user and captured by our query. The sample includes users located in the UK and Switzerland as control group and users located in Austria, France, Germany, Ireland, and Italy as treatment group. We use tweets from the period between 19^{th} February to 15^{th} March 2022. We estimate Equation 2 including user- and day-fixed effect, clustering standard errors at the user level, and controlling for word count, mentions count, and hashtags count.

speech rights, as well as questioning the effectiveness of such policies within democratic frameworks. This paper offers an empirical perspective on evaluating the impacts of media censorship within democratic contexts.

To investigate the effect of censorship in democracy, our study examines the consequences of an unprecedented decision taken by the European Union in March 2022, during the early stages of the Russo-Ukrainian conflict initiated by Russia's invasion of Ukraine. It quickly became evident that the conflict extended beyond physical confrontations, encompassing a digital dimension characterized by a surge of propaganda, misinformation, and biased narratives about the conflict flooding the European internet space. In response to this influx, European institutions enacted a ban on two of the principal channels of state-led propaganda, Russia Today and Sputnik, along with their affiliates. Leveraging the natural experiment emerging from the differential timing and geographical spread of the ban's implementation,

Table 5: *User-day level two-periods TWFE: Top 0.5% suppliers*

	(1) Avg. media slant	(2) % pro-Russia tweets	(3) % pro-Russia retweets	(4) Tot. Pro-Russia tweets	(5) Tot. pro-Russia retweets
EU × after-ban	0.094	0.146	0.092	0.256	0.259
	[0.179]	[0.078]	[0.075]	[1.205]	[0.508]
User FEs	yes	yes	yes	yes	yes
Day FEs	yes	yes	yes	yes	yes
Observations	640	561	541	640	640
R^2	0.473	0.293	0.261	0.460	0.407
Pre-period mean of DV	0.398	0.295	0.294	10.860	10.899
% of mean	23.63	49.53	31.24	2.36	2.38

Notes: The table displays coefficients from estimating Equation 2 to assess the ban's impact on the top 0.5% of *suppliers*. These *suppliers* are defined as users who posted at least one tweet or retweet with a slant over 1 prior to the ban and ranked in the top 0.5% for activity among all *suppliers* in the pre-ban period. Column 1 shows the effects on our measure of media slant, the intensive margin. Columns 2 and 3 show effects on our measure of the extensive margin, respectively, share of tweets and pro-Russia retweets. Columns 4 and 5 show the effect on the total number of pro-Russia tweets and retweets, respectively, produced by the author in the time period. The sample includes users in the UK and Switzerland as control group and users in Austria, France, Germany, Ireland, and Italy as treatment group. We use tweets from the period between 19^{th} February to 15^{th} March 2022. We estimate Equation 2 including user- and day-fixed effects, and controlling for word count, mentions count, and hashtags count. Standard errors are clustered at the user level are reported in brackets.

our analysis investigates the extent to which the prevalence and intensity of pro-Russia media slant were reduced in some of the countries enforcing the ban – Austria, France, Germany, Italy, and Ireland – compared to Switzerland and the United Kingdom, which did not adopt similar measures at the same time.

We find that the ban is associated with a reduction of the average pro-Russian slant of users who had previously engaged with the banned outlets, with a notable decrease of roughly 63.1% relative to the pre-ban mean in average slant. In turn, we find no meaningful effect on the extensive margin of slant measured as the share of tweets and retweets classified as pro-Russia slanted content. Moreover, our study suggests a short-lived effect of the ban. Further, we document indirect effects on users who did not directly interact with the outlets but find them to be limited in magnitude. This suggests that while the ban initially moderated pro-Russian slant within the directly affected subset of users, its broader and lasting effect on the overall discussion was limited.

Our study further explores potential mechanisms that might have offset the ban's intended effect. One plausible explanation for the ban's limited impact could be the emergence of new *suppliers* of slanted content into the media market after the ban. Although our analysis shows an increase in the share of users actively disseminating slanted content post-ban – likely in reaction to the escalating conflict – this increase was significantly more pronounced in countries not enforcing the ban, as opposed to those within the EU. Next, we explore an alternative explanation: users who were already disseminating pro-Russian content before the ban might have increased their activity to fill the gap left by Russia Today and Sputnik. Our findings indicate that this is a more probable route through which the ban's effects were counteracted. Specifically, we provide suggestive evidence that the most active *suppliers* in the EU prior to the ban increase their activity relative to their counterparts in non affected countries.

To be clear, the findings of this study should not be taken as an endorsement of banning media outlets nor as conclusive evidence of the effectiveness of these measures. Instead, we interpret our results

as suggestive evidence that regulating the media market using censorship in a democratic context can have an effect on online discourse. However, its effectiveness might be limited due to adjustments by consumers and producers of slanted content. As our study is limited to a short time window around the ban, the results should be interpreted with caution when evaluating the long-term effects or any overall welfare effects, considering the potential cost of using censorship as a policy tool.

References

Acemoglu, Daron, Tarek A. Hassan, and Ahmed Tahoun (2018). "The Power of the Street: Evidence from Egypt's Arab Spring". In: *The Review of Financial Studies* 31.1, pp. 1–42. ISSN: 0893-9454, 1465-7368.

Adena, Maja et al. (2015). "Radio and the Rise of The Nazis in Prewar Germany*". In: *The Quarterly Journal of Economics* 130.4, pp. 1885–1939. ISSN: 0033-5533, 1531-4650.

Allcott, Hunt and Matthew Gentzkow (2017). "Social Media and Fake News in the 2016 Election". In: *Journal of Economic Perspectives* 31.2, pp. 211–236. ISSN: 0895-3309.

Allcott, Hunt et al. (2020). "The Welfare Effects of Social Media". In: *American Economic Review* 110.3, pp. 629–676. ISSN: 0002-8282.

Baade, Björnstjern (2022). "The EU's "Ban" of RT and Sputnik: A Lawful Measure Against Propaganda for War". In: *Verfassungsblog*.

Badawy, Adam et al. (Dec. 2019). "Characterizing the 2016 Russian IRA influence campaign". In: *Social Network Analysis and Mining* 9.1, p. 31. ISSN: 1869-5450, 1869-5469.

Barbera, Pablo (2014). "How Social Media Reduces Mass Political Polarization. Evidence from Germany, Spain, and the US". In: *ob Market Paper, New York University* 46, pp. 1–46.

Becker, Sascha O, Francisco J Pino, and Jordi Vidal-Robert (2021). "Freedom of the Press? Catholic Censorship during the Counter-Reformation". In: *CEPR Discussion Paper No. DP16092*.

Bjørnskov, Christian and Stefan Voigt (2021). "Is Constitutionalized Media Freedom Only Window Dressing? Evidence from Terrorist Attacks". In: *Public Choice* 187.3, pp. 321–348. ISSN: 0048-5829, 1573-7101.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess (Jan. 16, 2024). *Revisiting Event Study Designs: Robust and Efficient Estimation*. arXiv: 2108.12419[econ].

Bursztyn, Leonardo et al. (2019). *Social Media and Xenophobia: Evidence from Russia*. w26567. Cambridge, MA: National Bureau of Economic Research, w26567.

Caetano, Carolina and Brantly Callaway (May 18, 2023). *Difference-in-Differences with Time-Varying Covariates in the Parallel Trends Assumption*. arXiv: 2202.02903[econ].

Cage, Julia, Nicolas Herve, and Beatrice Mazoyer (2020). "Social Media and Newsroom Production Decisions". In: *SSRN Electronic Journal*. ISSN: 1556-5068.

Callaway, Brantly and Pedro H.C. Sant'Anna (Dec. 2021). "Difference-in-Differences with multiple time periods". In: *Journal of Econometrics* 225.2, pp. 200–230. ISSN: 03044076.

Campante, Filipe, Ruben Durante, and Francesco Sobbrio (2018). "Politics 2.0: The Multifaceted Effect of Broadband Internet on Political Participation". In: *Journal of the European Economic Association* 16.4, pp. 1094–1136. ISSN: 1542-4766, 1542-4774.

Campante, Filipe, Ruben Durante, and Andrea Tesei, eds. (2023). *The Political Economy of Social Media*. Paris & London: CEPR Press. 211 pp.

Chandrasekharan, Eshwar et al. (2017). "You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech". In: *Proceedings of the ACM on Human-Computer Interaction* 1 (CSCW), pp. 1–22. ISSN: 2573-0142.

Chang, Neng-Chieh (May 1, 2020). "Double/debiased machine learning for difference-in-differences models". In: *The Econometrics Journal* 23.2, pp. 177–191. ISSN: 1368-4221, 1368-423X.

Chen, Yuyu and David Y. Yang (2019). "The Impact of Media Censorship: 1984 or Brave New World?" In: *American Economic Review* 109.6, pp. 2294–2332. ISSN: 0002-8282.

Chernozhukov, Victor et al. (Feb. 1, 2018). "Double/debiased machine learning for treatment and structural parameters". In: *The Econometrics Journal* 21.1, pp. C1–C68. ISSN: 1368-4221, 1368-423X.

Chu, Zi et al. (2012). "Detecting Automation of Twitter Accounts: Are You a Human, Bot, or Cyborg?" In: *IEEE Transactions on Dependable and Secure Computing* 9.6, pp. 811–824. ISSN: 1545-5971.

Corduneanu-Huci, Cristina and Alexander Hamilton (2022). "Selective Control: The Political Economy of Censorship". In: *Political Communication* 39.4, pp. 517–538. ISSN: 1058-4609.

Eady, Gregory et al. (Jan. 9, 2023). "Exposure to the Russian Internet Research Agency foreign influence campaign on Twitter in the 2016 US election and its relationship to attitudes and voting behavior". In: *Nature Communications* 14.1, p. 62. ISSN: 2041-1723.

Enikolopov, Ruben, Maria Petrova, and Ekaterina Zhuravskaya (2011). "Media and Political Persuasion: Evidence from Russia". In: *American Economic Review* 101.7, pp. 3253–3285. ISSN: 0002-8282.

Ershov, Daniel and Juan S Morales (2021). "Sharing News Left and Right: The Effects of Policies Targeting Misinformation on Social Media". In: *Carlo Alberto Notebooks No. 651*.

Gao, Tianyu, Xingcheng Yao, and Danqi Chen (May 18, 2022). *SimCSE: Simple Contrastive Learning of Sentence Embeddings*. arXiv: 2104.08821[cs].

Gehlbach, Scott and Konstantin Sonin (2014). "Government control of the media". In: *Journal of Public Economics* 118, pp. 163–171. ISSN: 00472727.

Gehring, Kai and Matteo Grigoletto (2023). "Analyzing Climate Change Policy Narratives with the Character-Role Narrative Framework". In: *SSRN Electronic Journal*. ISSN: 1556-5068.

Gennaro, Gloria and Elliott Ash (2023). "Emotion and Reason in Political Language". In: *The Economic Journal* 133.650, pp. 904–904. ISSN: 0013-0133, 1468-0297.

Gentzkow, Matthew and Jesse M Shapiro (2010). "What Drives Media Slant? Evidence From U.S. Daily Newspapers". In: *Econometrica* 78.1, pp. 35–71. ISSN: 0012-9682.

Gentzkow, Matthew and Jesse M. Shapiro (2011). "Ideological Segregation Online and Offline". In: *The Quarterly Journal of Economics* 126.4, pp. 1799–1839. ISSN: 0033-5533, 1531-4650.

Guriev, Sergei, Nikita Melnikov, and Ekaterina Zhuravskaya (2021). "3G Internet and Confidence in Government". In: *The Quarterly Journal of Economics* 136.4, pp. 2533–2613. ISSN: 0033-5533, 1531-4650.

Guriev, Sergei and Daniel Treisman (2022). *Spin Dictators: The Changing Face of Tyranny in the 21st Century*. 2022nd ed. Princeton University Press.

Halberstam, Yosh and Brian Knight (2016). "Homophily, Group Size, and the Diffusion of Political Information in Social Networks: Evidence from Twitter". In: *Journal of Public Economics* 143, pp. 73–88. ISSN: 00472727.

Jhaver, Shagun et al. (2021). "Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter". In: *Proceedings of the ACM on Human-Computer Interaction* 5 (CSCW2), pp. 1–30. ISSN: 2573-0142.

Jiménez-Durán, Rafael (2023). "The Economics of Content Moderation: Theory and Experimental Evidence from Hate Speech on Twitter". In: *SSRN Electronic Journal*. ISSN: 1556-5068.

Jiménez Durán, Rafael, Karsten Müller, and Carlo Schwarz (2022). "The Effect of Content Moderation on Online and Offline Hate: Evidence from Germany's NetzDG". In: *SSRN Electronic Journal*. ISSN: 1556-5068.

Kellam, Marisa and Elizabeth A. Stein (2016). "Silencing Critics: Why and How Presidents Restrict Media Freedom in Democracies". In: *Comparative Political Studies* 49.1, pp. 36–77. ISSN: 0010-4140, 1552-3829.

Levy, Ro'ee (2021). "Social Media, News Consumption, and Polarization: Evidence from a Field Experiment". In: *American Economic Review* 111.3, pp. 831–870. ISSN: 0002-8282.

Morales, Juan S. (2020). "Perceived Popularity and Online Political Dissent: Evidence from Twitter in Venezuela". In: *The International Journal of Press/Politics* 25.1, pp. 5–27. ISSN: 1940-1612, 1940-1620.

Mullainathan, Sendhil and Andrei Shleifer (2005). "The Market for News". In: *American Economic Review* 95.4, pp. 1031–1053. ISSN: 0002-8282.

Müller, Karsten and Carlo Schwarz (2022). "The Effects of Online Content Moderation: Evidence from President Trump's Account Deletion". In: *SSRN Electronic Journal*. ISSN: 1556-5068.

– (2023). "From Hashtag to Hate Crime: Twitter and Antiminority Sentiment". In: *American Economic Journal: Applied Economics* 15.3, pp. 270–312. ISSN: 1945-7782, 1945-7790.

Ni, Jianmo et al. (2021). "Sentence-T5: Scalable Sentence Encoders from Pre-trained Text-to-Text Models". In: Publisher: arXiv Version Number: 3.

Roth, Jonathan (2024). "Interpreting Event-Studies from Recent Difference-in-Differences Methods". In: Publisher: [object Object] Version Number: 1.

Shadmehr, Mehdi and Dan Bernhardt (2015). "State Censorship". In: *American Economic Journal: Microeconomics* 7.2, pp. 280–307. ISSN: 1945-7669, 1945-7685.

Simonov, Andrey and Justin Rao (2022). "Demand for Online News under Government Control: Evidence from Russia". In: *Journal of Political Economy* 130.2, pp. 259–309. ISSN: 0022-3808, 1537-534X.

Tabassum, Fatima et al. (2023). "How Many Features Do We Need to Identify Bots on Twitter?" In: *Information for a Better World: Normality, Virtuality, Physicality, Inclusivity*, pp. 312–327.

Yanagizawa-Drott, David (2014). "Propaganda and Conflict: Evidence from the Rwandan Genocide". In: *The Quarterly Journal of Economics* 129.4, pp. 1947–1994. ISSN: 0033-5533, 1531-4650.

Censorship in Democracy

Marcel Caesmann*, Janis Goldzycher°, Matteo Grigoletto**, Lorenz Gschwent°°

Appendix

*University of Zurich, e-mail: *marcel.caesmann@econ.uzh.ch*

°University of Zurich, e-mail: *goldzycher@cl.uzh.ch*

**University of Bern, Wyss Academy for Nature at the University of Bern, e-mail: *matteo.grigoletto@unibe.ch*

°°University of Duisburg-Essen, RTG Regional Disparities and Economic Policy, e-mail: *lorenz.gschwent@uni-due.de*

Appendix

Table of Contents

A Materials and Methods

In this Appendix, we offer additional information on the materials and methods used in our analysis. Specifically, we detail the process of data retrieval in Section A.1 and describe our method for user localization in Section A.2.

A.1 Twitter API

This section provides insights into the process of extraction and processing of data from Twitter. The download of tweets is done via the Twitter APIv2 that allows researchers to extract any tweets posted and not deleted in the platform since 2006, with a monthly cap of ten million tweets[10]. All the tweets we downloaded were posted between January 24^{th}, 2022, and April 4^{th}, 2022. Using the extracted data we create two datasets on data collected from Twitter: (1) a dataset of Ukrainian and Russian government-associated accounts and (2) a sample of tweets posted by users involved in the discussion about the unfolding conflict and invasion of Ukraine.

The first dataset consists of tweets posted by accounts affiliated with the Russian and Ukrainian governments. We provide a full list of the accounts in Table 2 in Section 3. We refer to the 5,993 Tweets from Russian government exponents and 9,451 Tweets from Ukrainian government exponents as *government tweets* (GT). This sample is key in constructing our measure of propaganda. The only filtering we do for this sample is on the language. To be part of our *government tweets* sample, a tweets has to come from the selected accounts, in the period of interest, and has to be in English language.

The second dataset consists of tweets posted by users involved in the discussion about the unfolding conflict and the later invasion of Ukraine. When extracting this data, a clear trade-off emerges. On one side, we want to ensure that we capture a representative sample of the conversation about the conflict. Hence, it is necessary to use a query that is not too restrictive. On the other side, we need to impose some restrictions to avoid false positives – tweets not primarily concerned with the conflict. Hence, our keyword query to solve this issue focuses on the main entities involved in the conflict: Russia, Ukraine, and NATO; this led to the following query: *russ* OR ukrain* OR nato OR otan*. We initially downloaded all tweets fulfilling these conditions posted between January 24^{th}, 2022, and April 4^{th}, 2022 and the following day-hours windows: 9 a.m. to 12 a.m., 3 p.m. to 6 p.m., and 8 p.m. to 11 p.m. This results in 7,865,321 extracted tweets by 1,942,979 users.

A.2 Geolocation

To create the final dataset comprising tweets from general Twitter users, we employ a geo-location process for user identification. It is crucial to remember that Twitter data acquired via the API does not automatically include geo-tag information. This means that while our query seeks tweets related to the conflict, these tweets could originate from users all around the world. To construct a dataset specifically from users in the EU, we proceed as follows. Our initial download resulted in data from 1,942,979 users. Utilizing the geo-location method outlined in Gehring and Grigoletto (2023), we then identify users located in our

[10] For more information see the Twitter Developer Platform documentation of the Search Tweets endpoint.

target countries: Austria, France, Germany, Ireland, Italy, Switzerland, and the UK, narrowing the group to 133,276 users. For these users, we subsequently download all English language tweets matching our query: *russ* OR ukrain* OR nato OR otan*. The emphasis on English tweets aligns with the language of the propaganda poles in our study. This process yields a dataset of 775,616 tweets, to which we refer to as the *user's tweets* (UT). For detailed information on the geo-location methodology, please see Gehring and Grigoletto (2023).

B Additional descriptive output

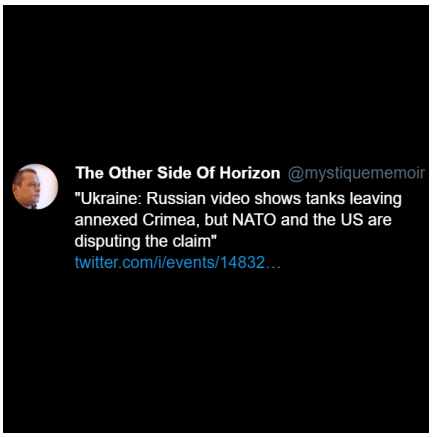
In this Appendix, we present additional descriptive data concerning the variables used in our analysis. Figure B.1 shows examples of tweets and their corresponding slant score.

Figure B.1: *Example tweets*

(a) *Pro-Ukrainian tweet*

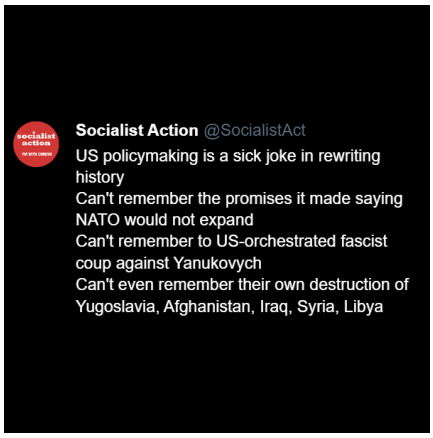
Index value: -1.95

(b) *Neutral tweet*



Index value: -0.01

(c) *Pro-Russian tweet*



Index value: 2.34

Notes: The Figure shows the examples of tweets and their corresponding slant measure. Back to Paper Section 3.

Figure B.2 depicts the time-series of the daily average levels of our slant measure for both EU and non-EU countries within our sample, focusing solely on original tweets. Conversely, Paper Figure 3 accounts for all tweet types in our dataset. A comparison of these figures reveals that the notable increase in the slant measure immediately preceding the ban, as observed in the latter figure, predominantly results from an increase in retweets. Such an increase is absent when only original tweets are considered.

Figure B.3 provides insights into the pro-Russia activity of users before the ban. In particular, Figure B.3a shows the average number of pro-Russia slanted tweets produced before the ban, by users that had no contacts with the banned outlets, and users that used to engage with the outlets, before the ban. Figure B.3b shows the same for *non-interaction* and *interaction* users before the ban, for retweets containing pro-Russia slanted content. These figures distinctly highlight the variance in pro-Russia content production between the two groups of users. Notably, users who previously engaged with the outlets produced three to four times more pro-Russia content than those who did not.

Figure B.4 delves into the activities of the most engaged pro-Russia users, specifically the top 0.5%,

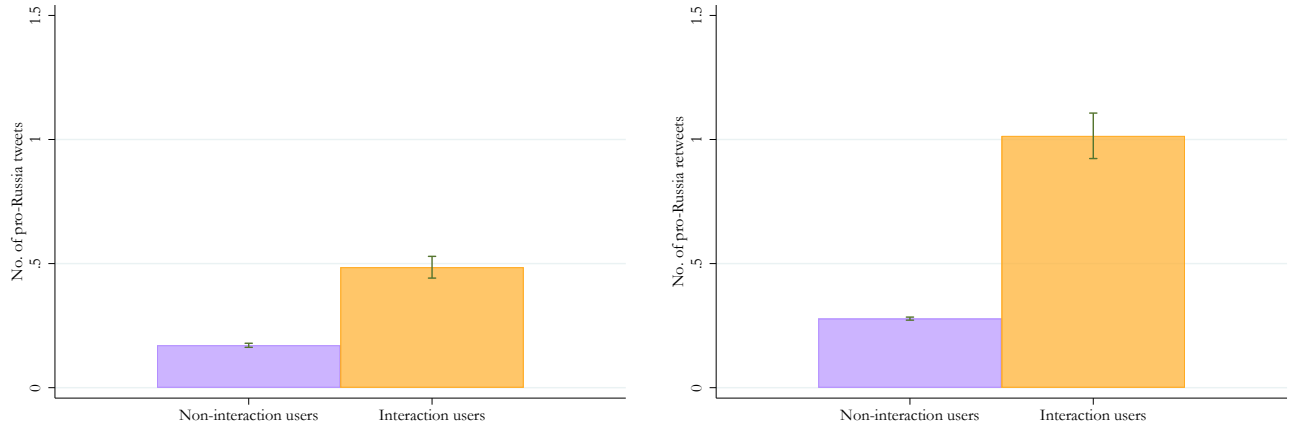
Figure B.2: *Time-series of average slant ratio in original tweets*

Notes: The figure shows the daily averages of our media slant measurement only in original tweets and excluding retweets, in the time-frame of our analysis, between 19^{th} February, 2022, to March 15^{th}, 2022. The measure is normalized to have a mean of 0 and a standard deviation of 1. When positive the measure indicates content closer to the Russian pole, and when negative it indicates content closer to the Ukrainian pole. In purple, we show the daily averages in the EU countries in our study, Austria, France, Germany, Ireland, and Italy, while in orange the daily averages in non-EU countries, United Kingdom and Switzerland. In grey, the difference between the two averages. We indicate in the graph the most relevant dates in our time period. Paper Figure 3 reproduces the same time-series but limiting the analysis to original tweets.

identified through a precise filtering process. Initially, we include all users who, before the ban, produced or shared any tweets or retweets exhibiting a pro-Russia slant score exceeding 1. From this pool, we examine the distribution of their activity levels prior to the ban. The top 0.5% represents users with the highest levels of pro-Russia slanted activity. The figure presents a comparison of average activities for these users before and after the ban, revealing that they are exceptionally active. Despite a noticeable decline in their activity following the ban, these users still produce, on average, almost 60 tweets per user, underscoring their significant engagement even in the face of restrictions.

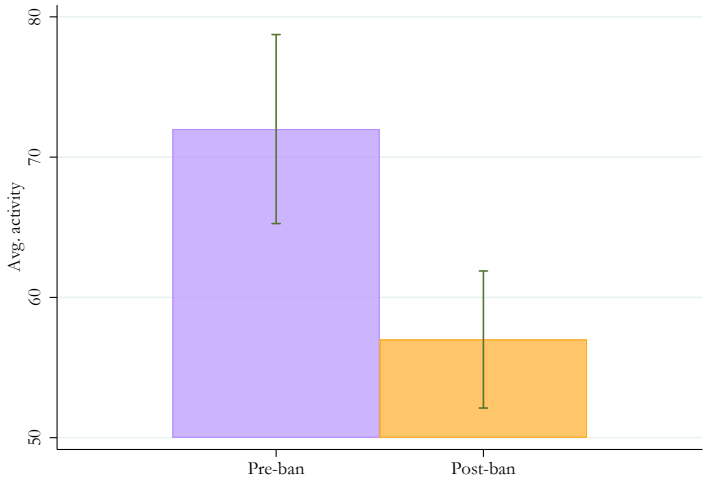
Figure B.3: *Pro-Russia activity of users before the ban*

(a) *Pro-Russia tweets* **(b)** *Pro-Russia retweets*



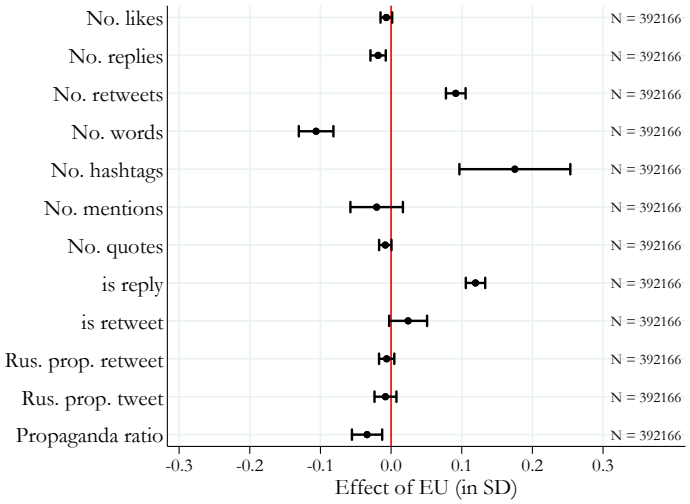
Notes: The Figure shows the average No. of pro-Russia tweets and the average No. of pro-Russia retweets – respectively sub-Figure B.3a and sub-Figure B.3b – of the *non-interaction users* vs. *interaction users* before the ban. Back to Paper Section 5.

Figure B.4: *Average activity of the top 0.5% suppliers*



Notes: The figure shows the average activity before and after the ban of the top 0.5% *suppliers* of pro-Russia slant. To define this *suppliers* we select all users that before the ban had produced or shared at least one tweet/retweet with slant above 1; among these suppliers, we take the distribution of total activity before the ban, and then select the top 0.5% most active users, comprising around 50 users. Back to Paper Section 5.

Figure B.5: *Balance in tweets characteristic before the ban: EU vs. non-EU countries*



Notes: The figure shows tests of balance on a number of pre-ban tweet features, between EU and non-EU countries in our sample. Back to Section 4.

C Robustness check: Alternative estimators

In this section, we delve into the rationale behind the use of some alternative estimators for the analysis. It is crucial to acknowledge that our research context presents significant challenges. Firstly, our dataset comprises a series of repeated cross-sections, not a standard panel. Secondly, while users may feature across multiple days, this results in an unbalance panel due to the extraction method. Thirdly, the users within our sample, despite originating from diverse locations, may be interconnected, adding another layer of complexity to our analysis.

As indicated in Equation 2, our analysis relies on a conditional parallel trends assumption, specifically conditioning on time-varying covariates. Caetano and Callaway (2023) discuss the challenges of applying traditional TWFE regressions in such contexts. Notably, TWFE regressions with time-varying covariates may encounter issues with negative weighting, similar to those identified in studies of staggered treatment designs. Moreover, TWFE regressions are vulnerable when parallel trends are influenced by the levels of time-varying covariates, rather than their changes. Consequently, we opt not to incorporate control variables in our TWFE regressions showcased in the paper.

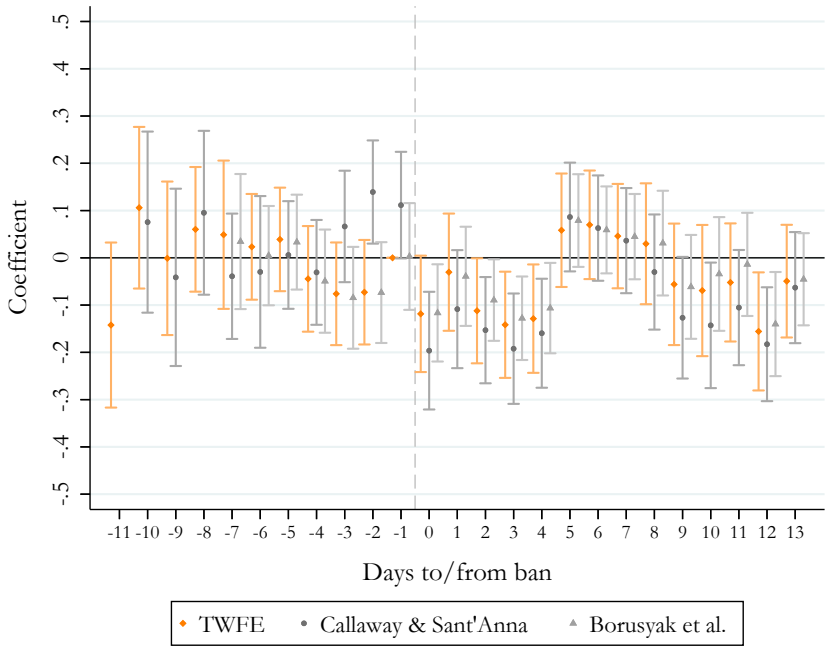
Caetano and Callaway (2023) demonstrate that "imputation estimators", akin to those proposed by Borusyak, Jaravel, and Spiess (2024), are capable of accurately estimating the treatment effect.[11] However, the existing implementations of these estimators typically compare outcomes for units observed just before the treatment, at $t-1$, with those post-treatment. Given our unbalanced panel and the absence of observations for all users right before the ban, this approach excludes a significant number of users. Conversely, the implementation by Callaway and Sant'Anna (2021) treats the data as repeated

[11] They further note that doubly-robust estimators, like those suggested by Chernozhukov et al. (2018); Chang (2020), are also applicable in such scenarios.

cross-sections, yet assumes covariates are time-invariant. Nevertheless, neither method perfectly suits our context.

Below, we show daily event studies displayed in the paper using two alternative estimators described above. For comparison we report on the graphs also estimates from a regression using OLS including user and day fixed effects. We are aware some recent work (see Roth (2024)) suggests to not display these coefficient together, nevertheless we do so for convenience of space and because they show a very similar pattern. Generally, the alternative estimators suggest very similar results and interpretation to what shown by the TWFE OLS estimation in the paper, despite presenting somewhat stronger coefficients.

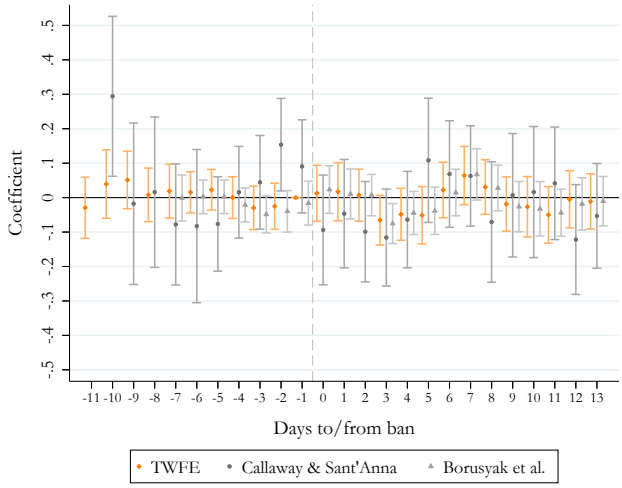
Figure C.1: *Daily event-study on our slant measure: Interaction users*



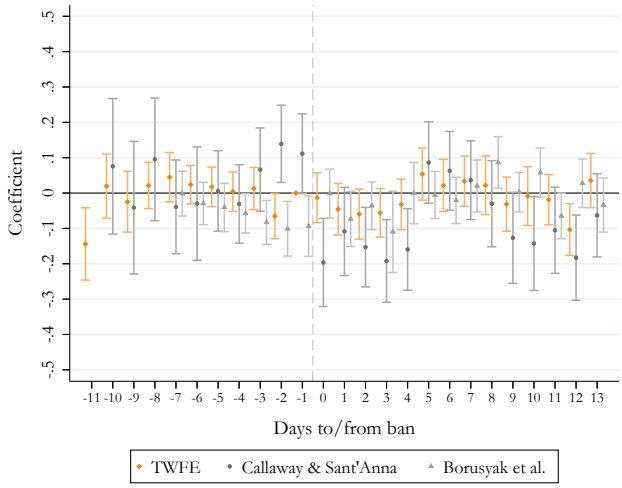
Notes: The figure displays coefficients and 95% confidence intervals from a daily event study, estimating regressions to assess the effect of the ban on our media slant measure, referred to as the intensive margin. The dependent variable is obtained by taking daily averages of media slant for each user that used to interact with Russia Today and Sputnik before the ban. In orange, we display regression results using a TWFE OLS estimator, while in grey, we present outcomes from the estimators proposed by Callaway and Sant'Anna (2021) and Borusyak, Jaravel, and Spiess (2024), which additionally incorporate time-varying control variables: word count, mentions count, and hashtags count. All models include the UK and Switzerland as control countries and Austria, France, Germany, Ireland, and Italy as treatment countries. They include tweets from the period between 19^{th} February to 15^{th} March 2022, they incorporate both user- and day-fixed effects and cluster standard errors at the user level. We control for word count, mentions count, and hashtags count. Back to Figure 4.

Figure C.2: *Daily event-study on share of slanted tweets and retweets: Interaction users*

(a) *Effect on pro-Russian slanted tweets*



(b) *Effect on pro-Russian slanted retweets*



Notes: The figure displays coefficients and 95% confidence intervals from a daily event study, estimating regressions to assess the effect of the ban on measures capturing the extensive margin of our analysis, for the *interaction users*. Figure 5a shows the impact on the share of daily tweets produced by a user and captured by our query, that can be classified as pro-Russia slant, hence having a media slant above 1. Figure 5b shows the same for retweets. In orange, we display regression results using a standard TWFE OLS estimation, while in grey, we present outcomes from the estimators proposed by Callaway and Sant'Anna (2021) and Borusyak, Jaravel, and Spiess (2024), which additionally incorporate time-varying control variables: word count, mentions count, and hashtags count. All models include the UK and Switzerland as control countries and Austria, France, Germany, Ireland, and Italy as treatment countries. They include tweets from the period between 19^{th} February to 15^{th} March 2022, they incorporate both user- and day-fixed effects and cluster standard errors at the user level. We control for word count, mentions count, and hashtags count. Back to Figure 5.

D Robustness check: Alternative text embedding models

In this Appendix, we provide alternative versions of the main findings from our study, utilizing different embedding models. Specifically, we explore several variations of Paper Figure 4. This figure originally investigates the effect of the ban on media slant within the *interaction users* group, where media slant is calculated by embedding user tweets through the sentence-t5-xl model. This involves computing the proximity ratio between tweets and the dynamically defined Russian and Ukrainian reference poles, which are updated daily based on data from the previous 14 days with a decay factor of 0.5. Here, we adjust various elements of this methodology to offer a broader perspective on our analysis.

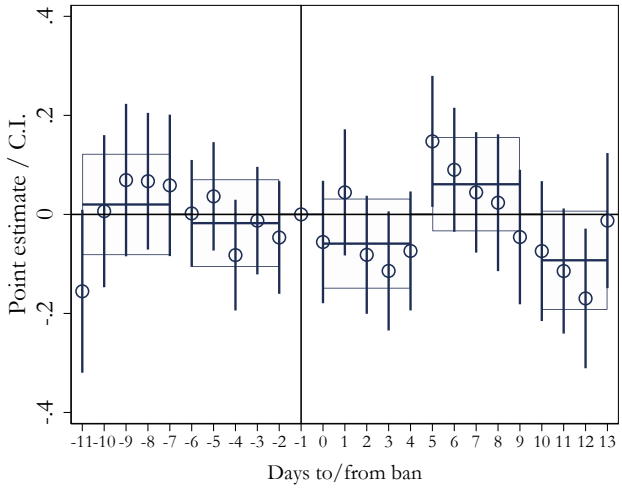
In Figure D.1 and D.2, we show the robustness of our results. We computed the measurement – in Figure D.1 Panel A – by only computing a single comparison pole for all Ukrainian and Russian government tweets over the full-time period (instead of 28 time-varying ones). The measure in Panel B of the same figure is similar to the one in Panel A but only considers government tweets that were posted before the ban. In Figure D.2 Panel A is the same as in the main part of the paper with an additional centered seven-day covering average smoothing. In Panel B of the same figure, we are substituting the vector representations from sentence-t5-xl with an alternative text embedding model, SimCSE, (Gao,

Yao, and Chen 2022). After sentence-t5-xl, SimCSE has shown the next best performances on various sentence embedding benchmarks. Therefore, we evaluated it as a suitable substitute.

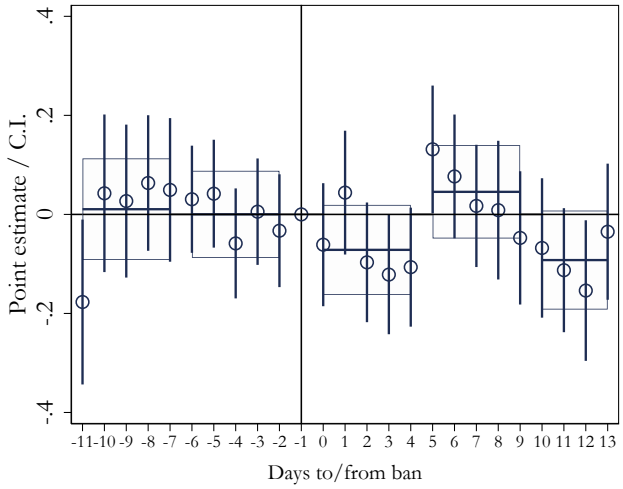
The results in the Figure D.1 show very similar patterns as in our main results, though the estimates after treatment are shifted upwards. We explain this by the static nature of these measures, which fail to capture the evolution of the government tweets as the full-scale war developed further. Nonetheless, the initial reduction in pro-Russian slant due to the ban is also captured by these measures. In Panel A of Figure D.2, results are hard to interpret given the strong pre-trends. Figure D.2 Panel B shows that the results also qualitatively hold for the alternative sentence embedding model, but with this model, pre-trends appear to be more problematic.

Figure D.1: *Daily event-study on our slant measure for interaction users: Alternative text embedding models (i.)*

(a) *Model: sentence-t5-xl static*



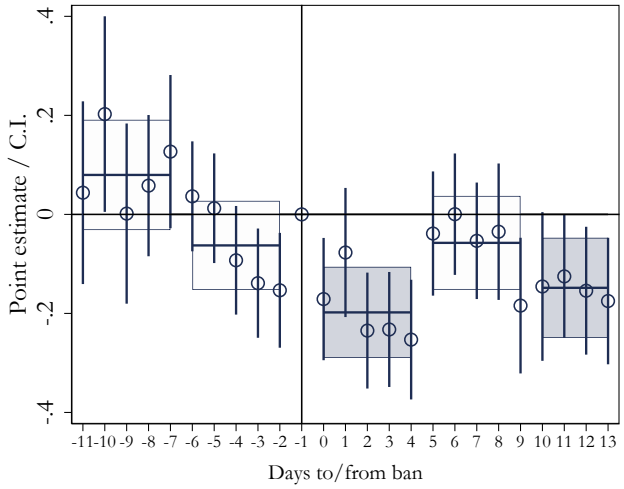
(b) *Model: sentence-t5-xl static pre-treatment*



Notes: The figure displays results reproducing what is shown in the paper Figure 4, but using alternative models to compute our media slant measure. Sub-figure D.1a shows results using a slant measure computed by only computing a single comparison pole for all Ukrainian and Russian government tweets over the full-time period. Sub-figure D.1b is similar but only considers government tweets that were posted before the ban.

Figure D.2: *Daily event-study on our slant measure for interaction users: Alternative text embedding models (ii.)*

(a) *Model: simCSE*



Notes: The figure displays results reproducing what is shown in the paper Figure 4, but using alternative models to compute our media slant measure. Sub-figure D.2a shows results substituting sentence-t5-xl with an alternative text embedding model, SimCSE (Gao, Yao, and Chen 2022).

E Robustness check: Excluding plausible bots

In this Appendix, we presents the findings from our analysis replicated from the main body of the paper, adjusting our sample with the removal of potential bots. To identify these bots, we rely on the criteria established by recent studies, which suggest that Twitter bots typically exhibit a high frequency of tweets per day (Tabassum et al. 2023) and a low 'reputation' ratio, calculated as the number of followers divided by the sum of the number of followers and the number of accounts followed (Chu et al. 2012). Following the recommendations of Gehring and Grigoletto (2023), we classify potential bots as accounts ranking in the upper 25% for daily tweet frequency and in the lower 25% for the reputation metric.

The exclusion of 32,432 tweets from 2,489 users identified by these criteria results in minimal changes to the distributions of descriptive statistics reported in Table E.1. Moreover, we demonstrate that the primary conclusions drawn in the main section of the paper remain consistent even after these accounts are omitted. In particular, we reproduce the results for *interaction users* excluding the potential bots. Figure E.1 shows the intensive margin analysis, Figure E.2 the extensive margin, Table E.2 the weekly interactions and Table E.3 the comparison with *non-interaction users*. An exception is represented by the heterogeneous effects of Figure E.3. It is seems the exclusion of bots, does change the effect on the most highly slanted *interaction users*, suggesting the bots population might be the most active in this sub-group.

Table E.1: *Summary statistics without plausible bots*

Panel A: Tweets

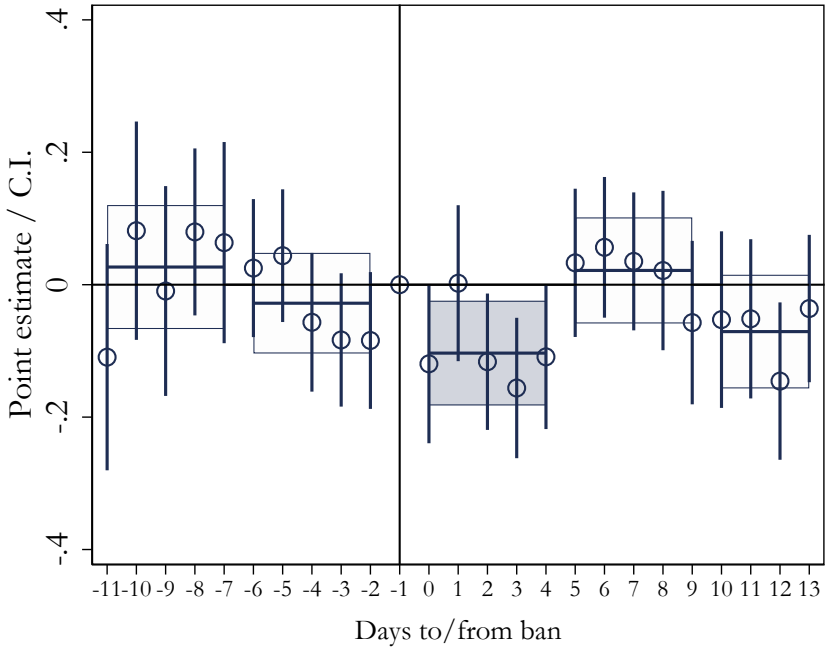
	Mean	Median	St. Dev.	Min.	Max.
Dependent Variables					
Propaganda ratio	-2.1e-10	.041	1	-4	4.8945765495
Russian propaganda tweet	.058	0	.23	0	1
Russian propaganda retweet	.1	0	.3	0	1
Tweet type					
Retweet	.53	1	.5	0	1
Reply	.088	0	.28	0	1
Tweet style					
numer of quotes of tweet	.12	0	5.8	0	2,369
No. of mentions	1.6	1	2.5	0	50
No. of hashtags	.44	0	1.6	0	42
No. of words	25	23	11	1	108
No. of Observations	762,332				

Panel B: Users

	Mean	Median	St. Dev.	Min.	Max.
User behavior					
No. tweets from user	2.7	1	12	0	1,543
No. retweets from user	3	1	11	0	676
No. replies from user	.51	0	2.1	0	209
No. russian propaganda tweets	.34	0	1.7	0	303
No. russian propaganda retweets	.59	0	2.3	0	151
Interacted with RT/Spk	.036	0	.19	0	1
No. retweets of RT/Spk	.00096	0	.041	0	5
Region					
European Union	.39	0	.49	0	1
No. of Observations	132,081				

Notes: Panel A reports descriptive statistics for the sample of tweets used in the analysis, posted by users that are not plausible bots and that we could locate in the countries of interest: Austria, France, Germany, Ireland, Italy, Switzerland and the United Kingdom. Tweets were extracted using the Historical Twitter APIv2, with the query: *ukrain* OR russ* OR NATO OR OTAN*, in the time window between February 19^{th} and March 15^{th} 2022. Panel B reports descriptive statistics on user characteristics, for users that posted tweets used in our analysis and described in Panel A. In both panels, for all variables we report mean, median, standard deviation, minimum, and maximum values. Paper Table 1 shows the same for the full sample.

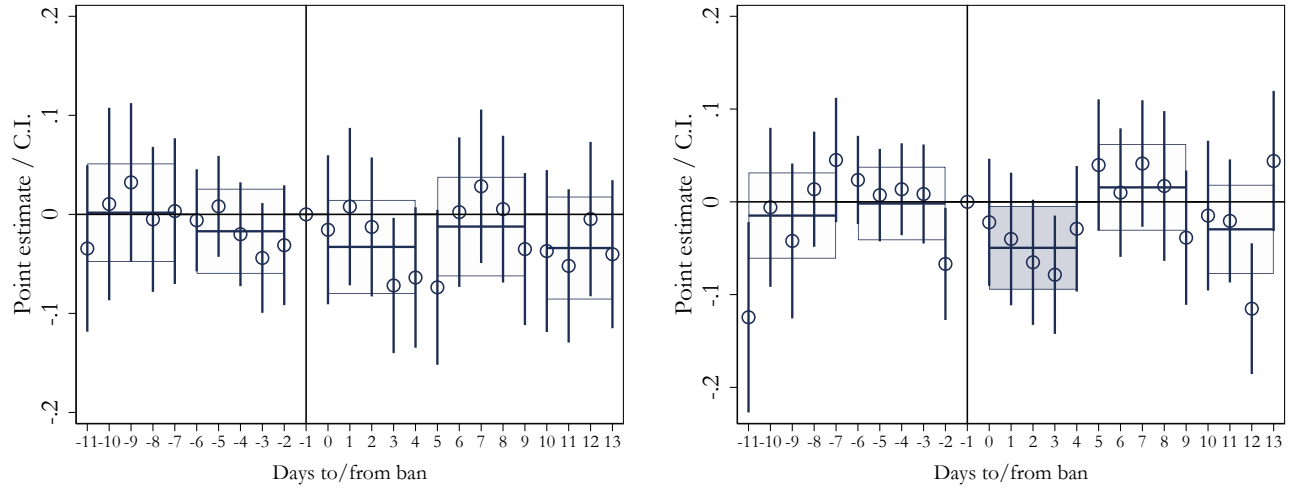
Figure E.1: *Daily event-study on share of slanted tweets and retweets: Interaction users excluding plausible bots*



Notes: The figure displays coefficients and 95% confidence intervals from estimating a daily event study version of Equation 2 excluding potential pots. The sample consists of *interaction users* – users who interacted with the banned outlets before the ban – and includes users located in the UK and Switzerland as control group and users located in Austria, France, Germany, Ireland, and Italy as treatment group. We exclude potential bots from the sample. Dependent variable is the daily average of slant in tweets for each user. We use tweets from the period between 19^{th} February to 15^{th} March 2022. We estimate Equation 2 including user- and day-fixed effects relative to the omitted day, 1^{st} March 2022 immediately before the implementation of the ban, and controlling for word count, mentions count, and hashtags count. In the aggregate specification, coefficients of interest are interactions between a dummy variable for aggregated intervals for 19^{th} to 23^{th} February, 24^{th} to 28 Februaryth, 2^{nd} to 6^{th} March, 7^{th} to 11^{th} March and 12^{th} to 15^{th} March, relative to the omitted day, 1^{st} March 2022 immediately before the implementation of the ban. Coefficient estimates on the day interactions are plotted as dots with their 95% confidence intervals indicated with vertical lines. Coefficient estimates on the aggregate interactions are shown with horizontal lines, and their 95% confidence intervals are indicated as boxes. We cluster standard errors at the user level. Paper Figure 4 shows results for all *interaction users*.

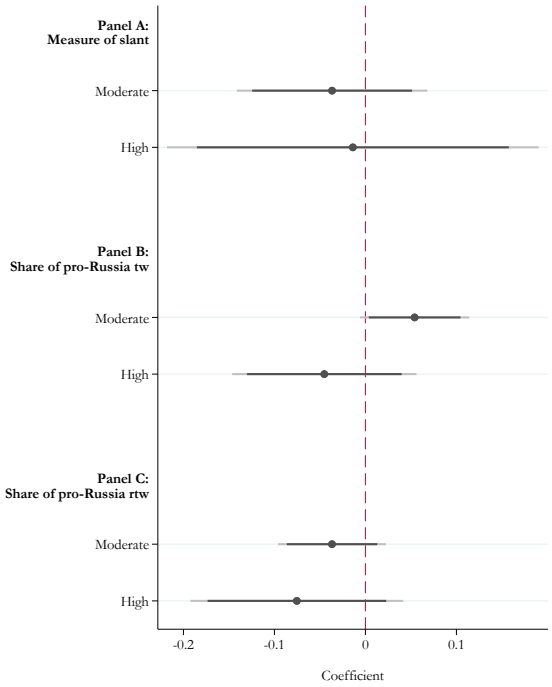
Figure E.2: *Daily event-study on share of slanted tweets and retweets: Interaction users excluding plausible bots*

(a) *Effect on pro-Russian slanted tweets* **(b)** *Effect on pro-Russian slanted retweets*



Notes: The figure displays coefficients and 95% confidence intervals from estimating a daily event study version of Equation 2 excluding potential bots. The sample consists of *interaction users* – users who interacted with the banned outlets before the ban – and includes users located in the UK and Switzerland as control group and users located in Austria, France, Germany, Ireland, and Italy as treatment group. We exclude potential bots from the sample. We use tweets from the period between 19^{th} February to 15^{th} March 2022. We estimate Equation 2 including user- and day-fixed effects relative to the omitted day, 1^{st} March 2022 immediately before the implementation of the ban, and controlling for word count, mentions count, and hashtags count. In the aggregate specification, coefficients of interest are interactions between a dummy variable for aggregated intervals for 19^{th} to 23^{th} February, 24^{th} to 28 Februaryth, 2^{nd} to 6^{th} March, 7^{th} to 11^{th} March and 12^{th} to 15^{th} March, relative to the omitted day, 1^{st} March 2022 immediately before the implementation of the ban. Coefficient estimates on the day interactions are plotted as dots with their 95% confidence intervals indicated with vertical lines. Coefficient estimates on the aggregate interactions are shown with horizontal lines, and their 95% confidence intervals are indicated as boxes. We cluster standard errors at the user level. Figure 5a display estimates using user's daily share of pro-Russian slanted tweets – defined as having a media slant measure above 1 – as dependent variable. Figure 5b display estimates using user's daily share of pro-Russian slanted retweets – defined as having a media slant measure above 1 – as dependent variable. Paper Figure 5 shows results for all *interaction users*.

Figure E.3: *Heterogeneous effects of the ban by pre-ban slant: Interaction users excluding plausible bots*



Notes: The figure presents coefficients from difference-in-differences regressions estimating Equation 2 and assessing the ban's heterogeneous impact on the *interaction users* that are not plausible bots. Users are divided into two groups: moderate, when the average slant in their tweets before the ban is in the bottom 75% of the distribution, and high, when the average slant in their tweets before the ban is in the top 25%. The coefficients are shown with 90% and 95% confidence intervals (95% in light grey). Panel A shows the results of our measure of media slant, the intensive margin, obtained by taking daily averages of media slant for each user that used to interact with Russia Today and Sputnik before the ban. Panel B and C show results on the daily proportion of tweets/retweets that can be classified as pro-Russia, out of all tweets/retweets produced by the user and captured by our query. All models include the UK and Switzerland as control countries and Austria, France, Germany, Ireland, and Italy as treatment countries. They include tweets from the period between 19^{th} February to 15^{th} March 2022, they incorporate both user- and day-fixed effects and cluster standard errors at the user level. We control for word count, mentions count, and hashtags count. Paper Figure 6 shows results for all *interaction users*.

Table E.2: *User-day level two-periods TWFE with post-ban weeks interactions: Interaction users excluding plausible bots*

	(1)	(2)	(3)	(4)	(5)
	Avg. media slant	% pro-Russia tweets	% pro-Russia retweets	Tot. Pro-Russia tweets	Tot. pro-Russia retweets
EU × 1st week after-ban	-0.052	-0.023	-0.022	-0.035	-0.051
	[0.023]	[0.014]	[0.014]	[0.022]	[0.027]
EU × 2nd week after-ban	-0.031	-0.008	-0.010	0.012	-0.023
	[0.025]	[0.016]	[0.014]	[0.021]	[0.024]
User FEs	yes	yes	yes	yes	yes
Day FEs	yes	yes	yes	yes	yes
Observations	29544	16508	19217	29544	29544
R^2	0.333	0.230	0.239	0.208	0.366
Pre-period mean of DV	-0.056	0.118	0.165	1.309	1.747
1st week % of mean	-91.74	-19.84	-13.59	-2.64	-2.92

Notes: The table displays coefficients from a two-period difference-in-differences regression analysis estimating Equation 2, examining the ban's differential impact in the two weeks following the ban, on the *interaction users* that are not plausible bots. Column 1 shows the effects on our measure of media slant, the intensive margin. Columns 2 and 3 show effects on our measure of the extensive margin, respectively share of tweets and retweets that are pro-Russia. Columns 4 and 5 show the effect on the total number of pro-Russia tweets and retweets, respectively, produced by the author in the time period. All models include the UK and Switzerland as control countries and Austria, France, Germany, Ireland, and Italy as treatment countries. They include tweets from the period between 19^{th} February to 15^{th} March 2022, they incorporate both user- and day-fixed effects and cluster standard errors at the user level. We control for word count, mentions count, and hashtags count. Paper Table 3 shows results including all *interaction users*.

Table E.3: *User-day level two-periods TWFE: Interaction and non-interaction users excluding potential bots*

Panel A: Consumers of content from outlets

	(1)	(2)	(3)	(4)	(5)
	Avg. media slant	% pro-Russia tweets	% pro-Russia retweets	Tot. Pro-Russia tweets	Tot. pro-Russia retweets
EU × after-ban	-0.043	-0.017	-0.017	-0.014	-0.039
	[0.021]	[0.012]	[0.012]	[0.018]	[0.022]
User FEs	yes	yes	yes	yes	yes
Day FEs	yes	yes	yes	yes	yes
Observations	29544	16508	19217	29544	29544
R^2	0.333	0.230	0.239	0.208	0.365
Pre-period mean of DV	-0.056	0.118	0.165	1.309	1.747
% of mean	-75.99	-14.22	-10.35	-1.08	-2.23

Panel A: Non-consumers of content from outlets

	(1)	(2)	(3)	(4)	(5)
	Avg. media slant	% pro-Russia tweets	% pro-Russia retweets	Tot. Pro-Russia tweets	Tot. pro-Russia retweets
EU × after-ban	-0.030	0.000	-0.037	-0.000	-0.012
	[0.008]	[0.004]	[0.004]	[0.005]	[0.005]
User FEs	yes	yes	yes	yes	yes
Day FEs	yes	yes	yes	yes	yes
Observations	312429	148704	178921	312429	312429
R^2	0.419	0.325	0.311	0.308	0.293
Pre-period mean of DV	-0.190	0.105	0.142	0.930	1.070
% of mean	-15.97	0.20	-25.96	-0.01	-1.16

Notes: The table displays coefficients from a two-period difference-in-differences regression analysis estimating Equation 2, examining the ban's impact on users who interacted with Russia Today and Sputnik before the ban in Panel A, and on user that had no interactions with the outlets in Panel B. We exclude in both panels users that are plausible bots. Column 1 shows the effects on our measure of media slant, the intensive margin. Columns 2 and 3 show effects on our measure of the extensive margin, respectively share of tweets and retweets that are pro-Russia. Columns 4 and 5 show the effect on the total number of pro-Russia tweets and retweets, respectively, produced by the author in the time period. All models include the UK and Switzerland as control countries and Austria, France, Germany, Ireland, and Italy as treatment countries. They include tweets from the period between 19^{th} February to 15^{th} March 2022, they incorporate both user- and day-fixed effects and cluster standard errors at the user level. We control for word count, mentions count, and hashtags count. Paper Table 4 shows results including also plausible bots.

F Robustness check: Without accounts created after the ban

In this Appendix, we presents the findings from our analysis replicated from the main body of the paper, adjusting our sample with the removal of accounts that were created after the ban from the sample. We do this to shed light on whether the creation of new accounts was an important mechanism of reaction to the ban. The exclusion of 2043 tweets from the 389 users that created their account after the ban, does not modify significantly the descriptive statistics and the results we presented in the main paper. Table F.1 shows the descriptive statistics. Figure F.1 shows the intensive margin analysis, Figure F.2 the extensive margin, Table F.2 the weekly interactions and Table F.3 the comparison with *non-interaction users*. An exception is represented by heterogeneous effects in Figure F.3. Despite the main Paper conclusions hold for the share of retweets, there seem to be no differentiale ffect for moderate and high slant for what concerns the intensive margin.

Table F.1: *Summary statistics without accounts created after the ban*

Panel A: Tweets

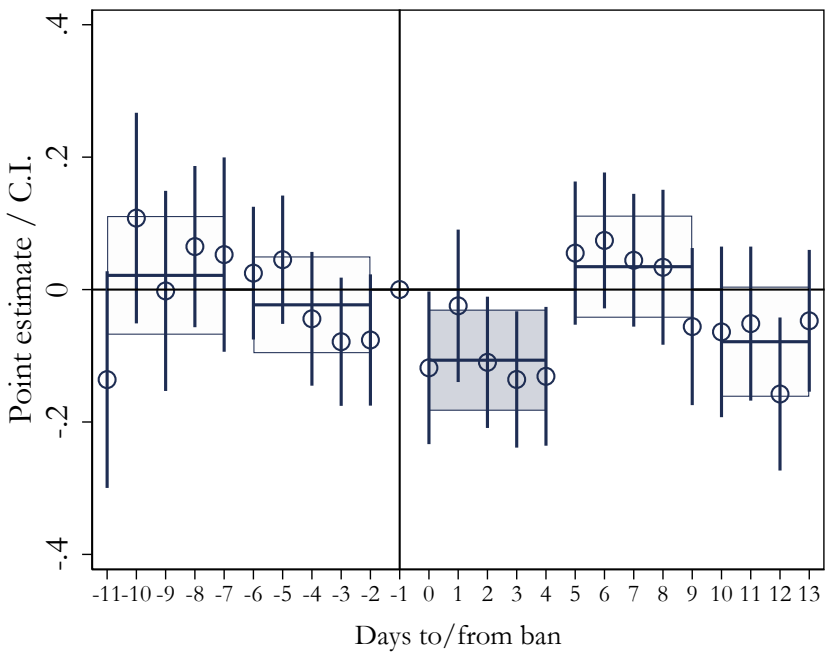
	Mean	Median	St. Dev.	Min.	Max.
Dependent Variables					
Propaganda ratio	-4.8e-10	.041	1	-4	4.8913640976
Russian propaganda tweet	.058	0	.23	0	1
Russian propaganda retweet	.1	0	.3	0	1
Tweet type					
Retweet	.53	1	.5	0	1
Reply	.09	0	.29	0	1
Tweet style					
numer of quotes of tweet	.11	0	5.7	0	2,369
No. of mentions	1.6	1	2.4	0	50
No. of hashtags	.45	0	1.6	0	42
No. of words	25	23	11	1	108
No. of Observations	792,721				

Panel B: Users

	Mean	Median	St. Dev.	Min.	Max.
User behavior					
No. tweets from user	2.8	1	12	0	1,543
No. retweets from user	3.1	1	11	0	676
No. replies from user	.53	0	2.2	0	209
No. russian propaganda tweets	.34	0	1.7	0	298
No. russian propaganda retweets	.6	0	2.3	0	151
Interacted with RT/Spk	.037	0	.19	0	1
No. retweets of RT/Spk	.001	0	.044	0	6
Region					
European Union	.39	0	.49	0	1
No. of Observations	134,181				

Notes: The table in Panel A reports descriptive statistics for the sample of tweets used in the analysis. The table in Panel B reports descriptive statistics on user characteristics, for users that posted tweets used in our analysis and described in Panel A. In both panels, for all variables we report mean, median, standard deviation, minimum, and maximum values. Paper Table 1 shows the same for the full sample.

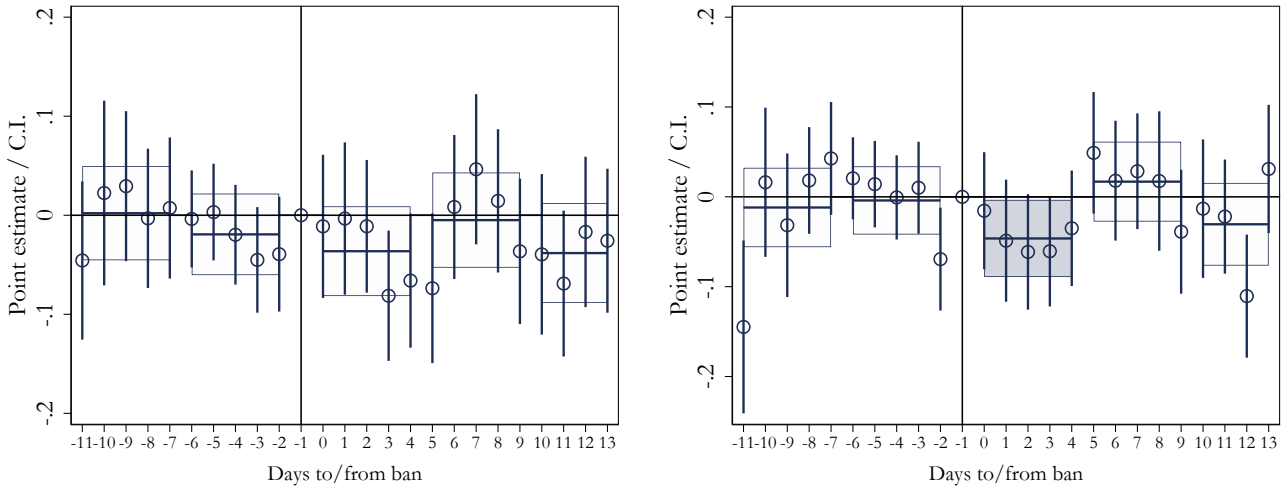
Figure F.1: *Daily event-study on share of slanted tweets and retweets: Interaction users excluding accounts created post-ban*



Notes: The figure displays coefficients and 95% confidence intervals from estimating a daily event study version of Equation 2 excluding accounts created post-ban. The sample consists of *interaction users* – users who interacted with the banned outlets before the ban – and includes users located in the UK and Switzerland as control group and users located in Austria, France, Germany, Ireland, and Italy as treatment group. We exclude all accounts created after the implementation of the ban. Dependent variable is the daily average of slant in tweets for each user. We use tweets from the period between 19^{th} February to 15^{th} March 2022. We estimate Equation 2 including user- and day-fixed effects relative to the omitted day, 1^{st} March 2022 immediately before the implementation of the ban. In the aggregate specification, coefficients of interest are interactions between a dummy variable for aggregated intervals for 19^{th} to 23^{th} February, 24^{th} to 28 Februaryth, 2^{nd} to 6^{th} March, 7^{th} to 11^{th} March and 12^{th} to 15^{th} March, relative to the omitted day, 1^{st} March 2022 immediately before the implementation of the ban. Coefficient estimates on the day interactions are plotted as dots with their 95% confidence intervals indicated with vertical lines. Coefficient estimates on the aggregate interactions are shown with horizontal lines, and their 95% confidence intervals are indicated as boxes. We cluster standard errors at the user level. We control for word count, mentions count, and hashtags count. Paper Figure 4 shows results for all *interaction users*.

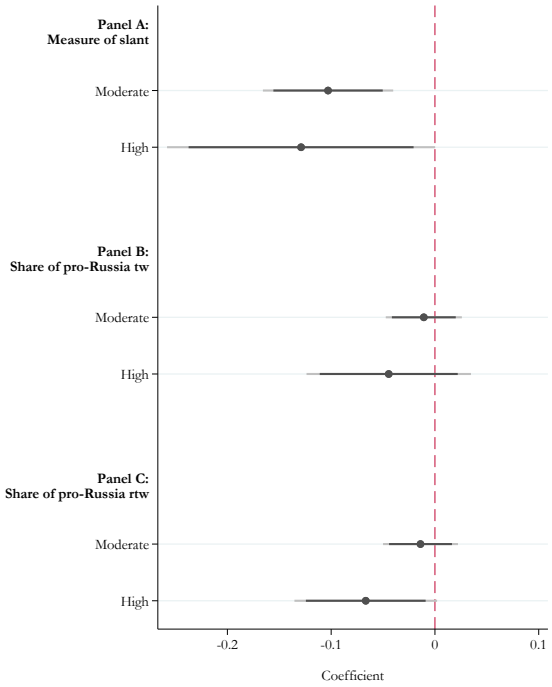
Figure F.2: *Daily event-study on share of slanted tweets and retweets: Interaction users excluding accounts created post-ban*

(a) *Effect on pro-Russian slanted tweets* **(b)** *Effect on pro-Russian slanted retweets*



Notes: The figure displays coefficients and 95% confidence intervals from estimating a daily event study version of Equation 2 excluding accounts created post-ban. The sample consists of *interaction users* – users who interacted with the banned outlets before the ban – and includes users located in the UK and Switzerland as control group and users located in Austria, France, Germany, Ireland, and Italy as treatment group. We exclude all accounts created after the implementation of the ban. We control for word count, mentions count, and hashtags count. We use tweets from the period between 19^{th} February to 15^{th} March 2022. We estimate Equation 2 including user- and day-fixed effects relative to the omitted day, 1^{st} March 2022 immediately before the implementation of the ban. In the aggregate specification, coefficients of interest are interactions between a dummy variable for aggregated intervals for 19^{th} to 23^{th} February, 24^{th} to 28 Februaryth, 2^{nd} to 6^{th} March, 7^{th} to 11^{th} March and 12^{th} to 15^{th} March, relative to the omitted day, 1^{st} March 2022 immediately before the implementation of the ban. Coefficient estimates on the day interactions are plotted as dots with their 95% confidence intervals indicated with vertical lines. Coefficient estimates on the aggregate interactions are shown with horizontal lines, and their 95% confidence intervals are indicated as boxes. We cluster standard errors at the user level. Figure 5a display estimates using user's daily share of pro-Russian slanted tweets – defined as having a media slant measure above 1 – as dependent variable. Figure 5b display estimates using user's daily share of pro-Russian slanted retweets – defined as having a media slant measure above 1 – as dependent variable. Paper Figure 5 shows results for all *interaction users*.

Figure F.3: *Heterogeneous effects of the ban by pre-ban level of pro-Russian slant: Interaction users excluding accounts created after the ban*



Notes: The figure presents coefficients from difference-in-differences regressions estimating Equation 2 and assessing the ban's heterogeneous impact on the *interaction users* that created their account any time before the ban. Users are divided into two groups: moderate, when the average slant in their tweets before the ban is in the bottom 75% of the distribution, and high, when the average slant in their tweets before the ban is in the top 25%. The coefficients are shown with 90% and 95% confidence intervals (95% in light grey). Panel A shows the results of our measure of media slant, the intensive margin, obtained by taking daily averages of media slant for each user that used to interact with Russia Today and Sputnik before the ban. Panel B and C show results on the daily proportion of tweets/retweets that can be classified as pro-Russia, out of all tweets/retweets produced by the user and captured by our query. All models include the UK and Switzerland as control countries and Austria, France, Germany, Ireland, and Italy as treatment countries. They include tweets from the period between 19^{th} February to 15^{th} March 2022, they incorporate both user- and day-fixed effects and cluster standard errors at the user level. We control for word count, mentions count, and hashtags count. Paper Figure 6 shows results for all *interaction users*.

Table F.2: *User-day level two-periods TWFE with post-ban weeks interactions: Interaction users excluding accounts created after the ban*

	(1)	(2)	(3)	(4)	(5)
	Avg. media slant	% pro-Russia tweets	% pro-Russia retweets	Tot. Pro-Russia tweets	Tot. pro-Russia retweets
EU × 1st week after-ban	-0.050	-0.024	-0.017	-0.037	-0.042
	[0.022]	[0.014]	[0.013]	[0.021]	[0.027]
EU × 2nd week after-ban	-0.034	-0.004	-0.012	0.013	-0.029
	[0.025]	[0.015]	[0.014]	[0.020]	[0.024]
User FEs	yes	yes	yes	yes	yes
Day FEs	yes	yes	yes	yes	yes
Observations	31076	17291	20436	31076	31076
R^2	0.332	0.229	0.239	0.207	0.364
Pre-period mean of DV	-0.058	0.118	0.164	1.307	1.799
1st week % of mean	-86.00	-20.54	-10.44	-2.80	-2.32

Notes: The table displays coefficients from a two-period difference-in-differences regression analysis estimating Equation 2, examining the ban's differential impact in the two weeks following the ban, on the *interaction users* that created their account any time before the ban. Column 1 shows the effects on our measure of media slant, the intensive margin. Columns 2 and 3 show effects on our measure of the extensive margin, respectively share of tweets and retweets that are pro-Russia. Columns 4 and 5 show the effect on the total number of pro-Russia tweets and retweets, respectively, produced by the author in the time period. All models include the UK and Switzerland as control countries and Austria, France, Germany, Ireland, and Italy as treatment countries. They include tweets from the period between 19^{th} February to 15^{th} March 2022, they incorporate both user- and day-fixed effects and cluster standard errors at the user level. We control for word count, mentions count, and hashtags count. Paper Table 3 shows results including all *interaction users*.

Table F.3: *User-day level two-periods TWFE: Interaction and non-interaction users excluding accounts created after the ban*

Panel A: Consumers of content from outlets

	(1)	(2)	(3)	(4)	(5)
	Avg. media slant	% pro-Russia tweets	% pro-Russia retweets	Tot. Pro-Russia tweets	Tot. pro-Russia retweets
EU × after-ban	-0.043	-0.016	-0.015	-0.015	-0.036
	[0.020]	[0.012]	[0.011]	[0.017]	[0.022]
User FEs	yes	yes	yes	yes	yes
Day FEs	yes	yes	yes	yes	yes
Observations	31076	17291	20436	31076	31076
R^2	0.332	0.229	0.239	0.207	0.364
Pre-period mean of DV	-0.058	0.118	0.164	1.307	1.799
% of mean	-74.10	-13.34	-9.17	-1.15	-2.01

Panel A: Non-consumers of content from outlets

	(1)	(2)	(3)	(4)	(5)
	Avg. media slant	% pro-Russia tweets	% pro-Russia retweets	Tot. Pro-Russia tweets	Tot. pro-Russia retweets
EU × after-ban	-0.030	0.001	-0.036	-0.004	-0.011
	[0.007]	[0.004]	[0.004]	[0.006]	[0.005]
User FEs	yes	yes	yes	yes	yes
Day FEs	yes	yes	yes	yes	yes
Observations	321772	152444	185939	321772	321772
R^2	0.417	0.324	0.308	0.293	0.291
Pre-period mean of DV	-0.189	0.105	0.142	0.932	1.089
% of mean	-15.81	0.70	-25.43	-0.47	-1.05

Notes: The table displays coefficients from a two-period difference-in-differences regression analysis estimating Equation 2, examining the ban's impact on users who interacted with Russia Today and Sputnik before the ban in Panel A, and on user that had no interactions with the outlets in Panel B. We exclude in both panels users that created their account after the ban. Column 1 shows the effects on our measure of media slant, the intensive margin. Columns 2 and 3 show effects on our measure of the extensive margin, respectively share of tweets and retweets that are pro-Russia. Columns 4 and 5 show the effect on the total number of pro-Russia tweets and retweets, respectively, produced by the author in the time period. All models include the UK and Switzerland as control countries and Austria, France, Germany, Ireland, and Italy as treatment countries. They include tweets from the period between 19^{th} February to 15^{th} March 2022, they incorporate both user- and day-fixed effects and cluster standard errors at the user level. We control for word count, mentions count, and hashtags count. Paper Table 4 shows results including also plausible bots.

G Robustness check: Different pro-Russia threshold

In this Appendix, we presents the findings from our analysis replicated from the main body of the paper, changing the threshold to define our binary variables of pro-Russia tweet and retweet. As a reminder, in the main analysis of the Paper, we use a threshold of 1 when we define our binary variables. This is to ensure we do not capture noise around the threshold of 0. Nevertheless, to prove additional robustness of our results we report in this appendix all results employing the binary variables, using a threshold of 0.

Table G.1: *Summary statistics using alternative threshold*

Panel A: Tweets

	Mean	Median	St. Dev.	Min.	Max.
Dependent Variables					
Propaganda ratio	-6.2e-10	.041	1	-4	4.8926959038
Russian propaganda tweet	.22	0	.42	0	1
Russian propaganda retweet	.29	0	.45	0	1
Tweet type					
Retweet	.53	1	.5	0	1
Reply	.09	0	.29	0	1
Tweet style					
numer of quotes of tweet	.11	0	5.7	0	2,369
No. of mentions	1.6	1	2.4	0	50
No. of hashtags	.45	0	1.6	0	42
No. of words	25	23	11	1	108
No. of Observations	794,764				

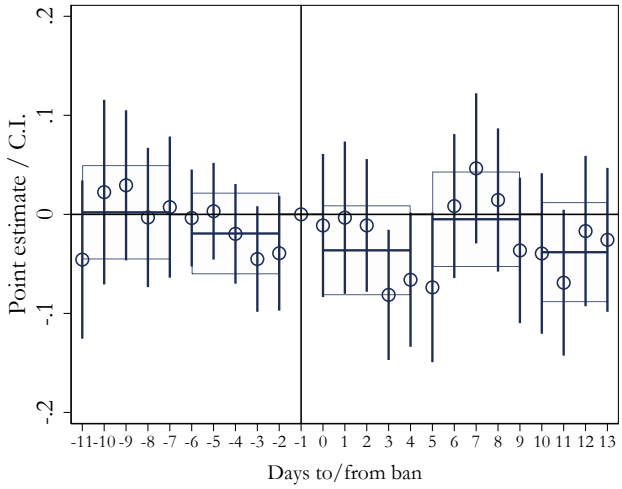
Panel B: Users

	Mean	Median	St. Dev.	Min.	Max.
User behavior					
No. tweets from user	2.8	1	12	0	1,543
No. retweets from user	3.1	1	11	0	676
No. replies from user	.53	0	2.2	0	209
No. russian propaganda tweets	1.3	0	5.5	0	712
No. russian propaganda retweets	1.7	0	6.3	0	403
Interacted with RT/Spk	.037	0	.19	0	1
No. retweets of RT/Spk	.001	0	.044	0	6
Region					
European Union	.39	0	.49	0	1
No. of Observations	134,570				

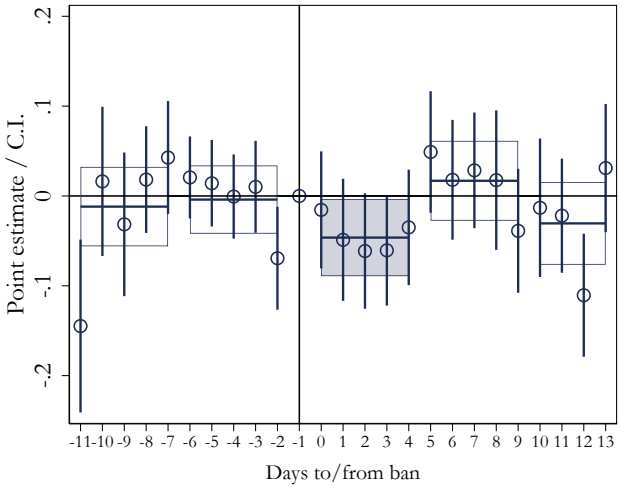
Notes: The table in Panel A reports descriptive statistics for the sample of tweets used in the analysis, posted by users that we could locate in the countries of interest: the United Kingdom and Switzerland, Austria, France, Germany, Ireland, and Italy. For binary variables we use the threshold 0 instead than 1 of our slant measure, for tweets and retweets to be defined as pro-Russia. Tweets were extracted using the Historical Twitter APIv2, with the query: *ukrain* OR russ* OR NATO OR OTAN*, in the time window between February 19^{th} and March 15^{th} 2022. The table in Panel B reports descriptive statistics on user characteristics, for users that posted tweets used in our analysis and described in Panel A. In both panels, for all variables we report mean, median, standard deviation, minimum, and maximum values. Paper Table 1 shows the same for the full sample.

Figure G.1: *Daily event-study on share of slanted tweets and retweets: Interaction users using alternative threshold*

(a) *Effect on pro-Russian slanted tweets*

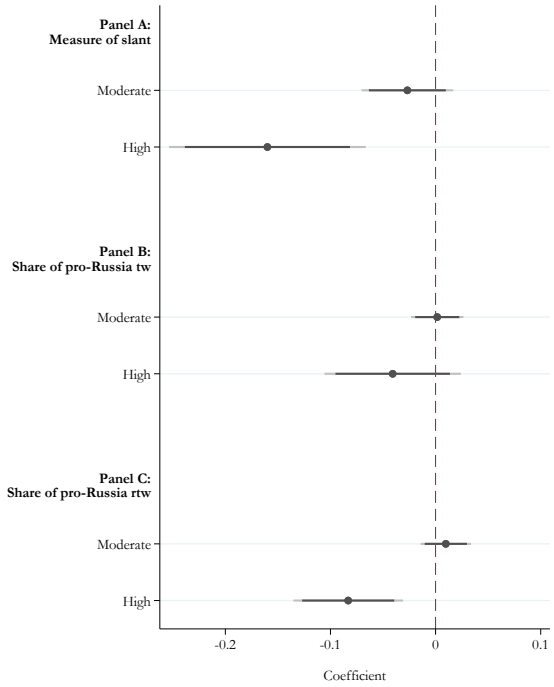


(b) *Effect on pro-Russian slanted retweets*



Notes: The figure displays coefficients and 95% confidence intervals from estimating a daily event study version of Equation 2 excluding accounts created post-ban. The sample consists of *interaction users* – users who interacted with the banned outlets before the ban – and includes users located in the UK and Switzerland as control group and users located in Austria, France, Germany, Ireland, and Italy as treatment group. We exclude all accounts created after the implementation of the ban. We use tweets from the period between 19^{th} February to 15^{th} March 2022. We estimate Equation 2 including user- and day-fixed effects relative to the omitted day, 1^{st} March 2022 immediately before the implementation of the ban. In the aggregate specification, coefficients of interest are interactions between a dummy variable for aggregated intervals for 19^{th} to 23^{th} February, 24^{th} to 28 Februaryth, 2^{nd} to 6^{th} March, 7^{th} to 11^{th} March and 12^{th} to 15^{th} March, relative to the omitted day, 1^{st} March 2022 immediately before the implementation of the ban. Coefficient estimates on the day interactions are plotted as dots with their 95% confidence intervals indicated with vertical lines. Coefficient estimates on the aggregate interactions are shown with horizontal lines, and their 95% confidence intervals are indicated as boxes. We cluster standard errors at the user level. We control for word count, mentions count, and hashtags count. Figure 5a display estimates using user's daily share of pro-Russian slanted tweets – defined as having a media slant measure above 0 – as dependent variable. Figure 5b display estimates using user's daily share of pro-Russian slanted retweets – defined as having a media slant measure above 0 – as dependent variable. Paper Figure 5 shows results for all *interaction users* using a threshold of 1 to define pro-Russian slanted tweets and retweets.

Figure G.2: *Heterogeneous effects of the ban by pre-ban level of pro-Russian slant: Interaction users using alternative threshold*



Notes: The figure presents coefficients from DiD regressions estimation Equation and assessing the ban's heterogeneous impact on the *interaction users*. For binary variables we use the threshold 0 instead than 1 of our slant measure, for tweets and retweets to be defined as pro-Russia. Users are divided into two groups: moderate, when the average slant in their tweets before the ban is in the bottom 75% of the distribution, and high, for top 25%. The coefficients are shown with 90% and 95% confidence intervals (95% in light grey). Panel A shows the results for the intensive margin. Panel B and C show results on the daily proportion of tweets/retweets that can be classified as pro-Russia. All models include the UK and Switzerland as control countries and Austria, France, Germany, Ireland, and Italy as treatment countries. They include tweets from the period between 19^{th} February to 15^{th} March 2022, they incorporate both user- and day-fixed effects and cluster standard errors at the user level. We control for word count, mentions count, and hashtags count. Paper Figure 6 shows results for all *interaction users*.

Table G.2: *User-day level two-periods TWFE with post-ban weeks interactions: Interaction users using alternative threshold*

	(1)	(2)	(3)	(4)	(5)
	Avg. media slant	% pro-Russia tweets	% pro-Russia retweets	Tot. Pro-Russia tweets	Tot. pro-Russia retweets
EU × 1st week after-ban	-0.050	-0.036	-0.003	-0.091	0.014
	[0.022]	[0.017]	[0.015]	[0.043]	[0.046]
EU × 2nd week after-ban	-0.034	-0.005	0.029	0.044	0.060
	[0.025]	[0.020]	[0.017]	[0.051]	[0.054]
User FEs	yes	yes	yes	yes	yes
Day FEs	yes	yes	yes	yes	yes
Observations	31082	17297	20436	31082	31082
R^2	0.332	0.331	0.254	0.337	0.506
Pre-period mean of DV	-0.058	0.459	0.518	1.307	1.799
1st week % of mean	-87.08	-7.76	-0.55	-6.97	0.76

Notes: The table displays coefficients from a two-period difference-in-differences regression analysis estimating Equation 2, examining the ban's differential impact in the two weeks following the ban, on the *interaction users*. For binary variables we use the threshold 0 instead than 1 of our slant measure, for tweets and retweets to be defined as pro-Russia. Column 1 shows the effects on our measure of media slant, the intensive margin. Columns 2 and 3 show effects on our measure of the extensive margin, respectively share of tweets and retweets that are pro-Russia. Columns 4 and 5 show the effect on the total number of pro-Russia tweets and retweets, respectively, produced by the author in the time period. All models include the UK and Switzerland as control countries and Austria, France, Germany, Ireland, and Italy as treatment countries. They include tweets from the period between 19^{th} February to 15^{th} March 2022, they incorporate both user- and day-fixed effects and cluster standard errors at the user level. We control for word count, mentions count, and hashtags count. Paper Table 3 shows results including all *interaction users*.

Table G.3: *User-day level two-periods TWFE: Interaction and non-interaction users using alternative threshold*

Panel A: Consumers of content from outlets

	(1)	(2)	(3)	(4)	(5)
	Avg. media slant	% pro-Russia tweets	% pro-Russia retweets	Tot. Pro-Russia tweets	Tot. pro-Russia retweets
EU × after-ban	-0.043	-0.023	0.011	-0.032	0.034
	[0.020]	[0.016]	[0.013]	[0.041]	[0.043]
User FEs	yes	yes	yes	yes	yes
Day FEs	yes	yes	yes	yes	yes
Observations	31082	17297	20436	31082	31082
R^2	0.332	0.331	0.254	0.336	0.506
Pre-period mean of DV	-0.058	0.459	0.518	1.307	1.799
% of mean	-74.93	-4.93	2.15	-2.47	1.89

Panel A: Non-consumers of content from outlets

	(1)	(2)	(3)	(4)	(5)
	Avg. media slant	% pro-Russia tweets	% pro-Russia retweets	Tot. Pro-Russia tweets	Tot. pro-Russia retweets
EU × after-ban	-0.030	0.003	-0.006	-0.010	0.055
	[0.007]	[0.006]	[0.005]	[0.013]	[0.009]
User FEs	yes	yes	yes	yes	yes
Day FEs	yes	yes	yes	yes	yes
Observations	322338	152855	186148	322338	322338
R^2	0.417	0.409	0.325	0.293	0.409
Pre-period mean of DV	-0.189	0.404	0.477	0.932	1.089
% of mean	-15.87	0.83	-1.36	-1.12	5.06

Notes: The table displays coefficients from a two-period difference-in-differences regression analysis employing TWFE OLS estimator, examining the ban's impact on users who interacted with Russia Today and Sputnik before the ban in Panel A, and on user that had no interactions with the outlets in Panel B. For binary variables we use the threshold 0 instead than 1 of our slant measure, for tweets and retweets to be defined as pro-Russia. Column 1 shows the effects on our measure of media slant, the intensive margin. Columns 2 and 3 show effects on our measure of the extensive margin, respectively share of tweets and retweets that are pro-Russia. Columns 4 and 5 show the effect on the total number of pro-Russia tweets and retweets, respectively, produced by the author in the time period. All models include the UK and Switzerland as control countries and Austria, France, Germany, Ireland, and Italy as treatment countries. They include tweets from the period between 19^{th} February to 15^{th} March 2022, they incorporate both user- and day-fixed effects and cluster standard errors at the user level. We control for word count, mentions count, and hashtags count. Paper Table 4 shows results including also plausible bots.