

Adding Fuel to the (Gun)Fire: How Politicians Polarize the Public Debate*

Lehan Zhang¹ Gabriele Gratton² Pauline Grosjean² Hasin Yousaf³ [†]

October 15, 2026

Abstract

Do politicians polarize public debates? We collect 4.75 million tweets debating 57 U.S. mass shootings. We exploit variation in the timing of politicians' first interventions into each debate and show that partisan and issue-polarized language increases after a politician first tweets about the event. These changes are not preceded by trends in polarization or debate activity, and the timing of political interventions is unrelated to event characteristics or the timing or content of tweets from other celebrities or news organizations. Comparable interventions by non-political celebrities generate no such effects. Politicians' interventions are themselves unusually partisan, and interventions supplying partisan language produce larger subsequent shifts in both partisan and issue-polarized language. Our findings suggest that politicians supply partisan frames that shape subsequent public discourse and point to elite communication as a supply-side source of polarization on social media.

Keywords Twitter, Partisanship, Polarization, Rhetorical supply, Mass shooting

JEL codes D72, L82, P00, Z13

*We are grateful to Eleonora Alabrese, Elliott Ash, Milena Djourelova, Karsten Donnay, Ruben Durante, Paola Giuliano, Sophie Hatte, Federica Izzo, Federico Masera, Gerard Padró i Miquel, Ricardo Perez-Truglia, Francesco Sobbrío, Milan Svolik, Clémence Tricaud, Adam Zelizer, Katia Zhuravskaya, and audiences at the 3rd CEPR Workshop on Media, Technology, Politics, and Society, the Australian Political Economy Workshop, the 9th Economics of Media Bias Workshop, CESifo Digital Platforms Workshop, the 2023 Econometric Society Australasian Meeting, ENS Lyon, the 2025 Conference of the Society for Institutional and Organizational Economics, the 2026 Benelux PECO, University of Western Australia, Zurich Political Economy Seminar Series, and Tilburg Workshop on Economic Governance of Social Media for insightful comments and discussions. We thank Henrik Osterberg and Yang Wu for excellent research assistance. Gratton and Grosjean are recipients of Australian Research Council Future Fellowships (FT210100176 and FT190100298). Gabriele Gratton's research was supported under the Australian Research Council's Discovery Projects funding scheme (Project DP240103257). This project received IRB approval from UNSW (approval number HC220333).

[†]ETH Zurich. Email: lehan.zhang@gess.ethz.ch. ²UNSW and CEPR. Emails: g.gratton@unsw.edu.au, p.grosjean@unsw.edu.au. ³UNSW. Email: h.yousaf@unsw.edu.au.

1 Introduction

Public debate in the United States has become increasingly divisive and partisan. Policy-relevant events are quickly interpreted and described through partisan and polarized lenses, curtailing constructive debate (Abramowitz and Webster, 2016; Boxell et al., 2024; Djourelova et al., 2024; Yousaf, 2021) and profoundly reshaping beliefs and behavior (Bursztyn et al., 2020, 2025). To explain this rise of polarization, a growing literature has pointed to the contemporary rise of social media and their role in amplifying political divisions and undermining social cohesion (see Campante et al., 2023; Zhuravskaya et al., 2020, for reviews).¹ To date, this literature has emphasized how social media foster demand-side mechanisms that fuel partisanship and polarization, including echo chambers, anonymity, and the consumption of fake news (see, e.g., Ederer et al., 2025; Guriev et al., 2023; Levy, 2021).

We instead focus on a supply-side effect of social media: enabling politicians’ direct communication.² Using Twitter debates surrounding major news events, we show that politicians’ direct communication on social media systematically steers online public debates towards more partisan and polarized language, whereas posts by other celebrities and news organizations do not produce the same shift. Moreover, the response is larger when the politicians’ posts *supply* partisan language, consistent with the interpretation that politicians supply partisan and polarized frames that shape subsequent discourse.

To conduct our analysis, we collect data on the universe of 4.75 million tweets related to the 57 mass shooting events covered on the front page of the *New York Times* between March 2016 (when Twitter rolled out its algorithmic timeline) and October 2022 (when Elon Musk acquired the company). We focus on mass shootings for several reasons. First, mass shootings are particularly salient news events, which have been shown to polarize lawmakers and voters on a wide range of political issues (Barilari, 2026; Yousaf, 2021). Second, mass shootings have a well-identified start time, making it possible to pin down exactly when the debate begins online. Third, their unanticipated nature ensures that politicians or political institutions did not influence the onset of the debate.

We extract all the tweets related to each event in the seven days after the event, thereby constructing a longitudinal “public debate” dataset.³ We define the first political interven-

¹Consequences can be tragic when online hate fuels physical violence and hate crimes (Bursztyn et al., 2023; Müller and Schwarz, 2021, 2023).

²Politicians are a natural source of polarizing frames. They are highly partisan and polarized (Barber and McCarty, 2015; Fowler et al., 2023), and many have more than one million Twitter followers, making them salient “influencers” who can amplify news and shape narratives (Boffa et al., 2024).

³A public debate includes, on average, 83,308 related tweets. As discussed in Section 2, the volume of tweets spikes in the first hour and a half after the start of an event and remains very stable after three days, so that our choice of stopping the data collection seven days after the event is innocuous.

tion as the first event-related post by a U.S. politician with at least one million followers.⁴ We also track interventions by celebrities who are not politicians, identified from a list of top 100 Twitter accounts by number of followers.⁵

For every tweet, we use dictionary-based methods to measure the tweet’s *partisanship* (Gentzkow et al., 2019) and *issue polarization* (Simchon et al., 2022).⁶ In addition, we show that our results are robust to using a complementary model-based measure of issue polarization, constructed by fine-tuning a language-model classifier on LLM-generated annotations. This measure also allows us to distinguish the ideological direction and issue domain of polarized discourse.

Our first contribution is to describe when and how politicians intervene in public debates. We uncover substantial variation in the identity and partisanship of the politicians who intervene in the public debate. In total, 67 (out of 70) different politicians intervene in 52 of the 57 public debates, with 29 distinct politicians intervening first in a debate—what we call the *first political intervention* in a public debate.

There is also substantial variation in the timing of political interventions and first political interventions, both with respect to the characteristics of the events (e.g., location, victim count, race of the shooter) and their unfolding (e.g., start, end, police intervention), and the timing of interventions by other politicians, news media, and other influencers. First political interventions occur, in general, well after interventions by news media organizations, but without any systematic pattern. Politicians also do not systematically follow other influencers, who tend to intervene later on. In addition, politicians do not systematically time their intervention with respect to key debate characteristics, such as the peak in the volume of related tweets, the popularity of users engaged in the debate, or the virality of tweets. In fact, even the first political intervention tends to arrive well after the debate has peaked.

Our second contribution is to document whether and how politicians polarize the public debate. Our two-way fixed effects and event-study analyses examine the shares of partisan and issue-polarized content in a debate after the first political intervention compared to before the intervention and compared to other debates in the same time

⁴Overall, 70 politicians have more than one million followers at the time of our study and are included in the analysis. Our results do not change if we focus on the much larger sets of the 117 U.S. politicians with more than 500,000 followers or the 202 U.S. politicians with more than 200,000 followers.

⁵See Section 2.5 for details.

⁶Gentzkow et al. (2019) measure partisanship from Congressional speech; Simchon et al. (2022)’s issue polarization dictionary is meant to capture “ideological, policy-based political divide.” We discuss in Section 5.5 how political interventions affect social tribalism, as captured by words that express loyalty and betrayal, as defined by the Moral Foundations Dictionary (Enke, 2020; Graham et al., 2009) and Simchon et al. (2022)’s affective polarization, capturing “dislike, distrust, and general animosity.”

interval since their respective onsets. Event-specific fixed effects account for any potential heterogeneity across events, which may be systematically correlated with the polarizing nature of the event. Because the nature of the debate may also naturally change over time as the debate unfolds and emotions get heated up or die off, we control for a set of fixed effects for highly granular time intervals since the onset of the debate,⁷ as well as for a time trend since the onset of the debate.⁸

A key concern in interpreting our results is that politicians may be responding to recent changes in the debate or recently released information. In that case, our estimations would measure the effect of these changes on subsequent tweets. However, our event-study results show that political interventions do not follow changes in volume or content (including partisan or issue-polarized) of the debate and they do not coincide with changes in the volume of tweets or the observable characteristics of the users who engage in the debate. Furthermore, political interventions do not coincide with news shocks: media coverage does not spike around political interventions, and news content released just before political interventions is not more likely to reveal particular new information about the event (e.g., on casualties, the shooter, or their motive). Finally, our estimates are unchanged when we control for media activity. Combined with the descriptive evidence on the timing of interventions, these results mitigate concerns that politicians strategically time their interventions in response to observable event characteristics or changes in its content or intensity.

We find that mass shooting events become partisan and issue-polarized topics *after* politicians intervene in the public debate. Political interventions trigger immediate and persistent increases in partisanship and in issue polarization. Within one hour of the first political intervention, the probability that a tweet contains partisan language is 43% greater than in the hour before the intervention (1.55% post vs. 1.08% pre-intervention). By two hours after the intervention, it is 88% greater. Issue polarization increases by, respectively, 17% and 40% relative to the pre-intervention mean, one and two hours after the first intervention.

Additional political interventions generate further, though diminishing, increases in partisan and issue-polarized language. LLM-based annotations confirm our results and further show that the discussion shifts toward gun policy, while issue polarization intensifies the most on *other* topics such as terrorism, media criticism, and race, thus politicizing

⁷Our baseline specification accounts for 30 minutes fixed effects, and we show that our results are robust to shorter or longer intervals.

⁸Our setting is not suitable for the use of calendar time fixed effects, since events occur at different dates with little overlap in the 7-day time frame after an event. In any case, calendar time fixed effects would be subsumed in our 30-minute time interval fixed effects.

the public debate beyond the gun policy debate.

We further show that political interventions increase tribal language, although they do not affect levels of affective polarization or hate speech. We also illustrate a sharp contrast between interventions by politicians vs. celebrities. Interventions by celebrities do not change the partisan or issue-polarized content of the public debate. If anything, they reduce tribalism and affective polarization, and do not influence hate speech.

To uncover potential mechanisms underlying these effects, we first document that the rhetoric of political interventions substantially differs from that of other social media users. In particular, interventions by politicians are uniquely partisan and issue-polarized. The first political intervention contains partisan language in 19% of cases, compared with 1.2% percent of pre-intervention tweets by regular users, making it about 16 times more likely to contain partisan language. Considering all interventions, politicians' tweets are approximately 17 percentage points more likely to contain partisan language and 19 percentage points more likely to contain issue-polarized language than interventions by celebrities or news media.

These differences in the rhetorical supply are crucial to understanding the polarizing effect of political interventions. Political interventions that are partisan are associated with a substantially larger increase in the subsequent partisanship and issue polarization of the debate. Further results show that these shifts are primarily driven by the content of tweets posted by newcomers to the debate (users who only join the specific debate after the first political intervention), and are stronger for newcomers. The increase is also larger among and those who are more politicized (more clearly identify with a political party or cause in their profile), although users without such identifiers shift too. By contrast, political interventions do not attract more users to the debate, nor do they change the composition of users in the debate (i.e., they do not mobilize more politicized users). Finally, while interventions by Democrat politicians are followed by an increase in Democratic-leaning partisan phrases only, interventions by Republicans equally spur Democratic- and Republican-leaning partisan phrases. Overall, these results are consistent with the view that political interventions inject partisan rhetoric in the public debate, placing news events under a partisan and polarized lens that reshapes the ensuing debate, particularly among those who only join the debate after the intervention.

Our findings contribute to three strands of research. First, we contribute to the literature on the political effects of social media. Platform access and exposure can affect political knowledge and polarization ([Allcott et al., 2020](#); [Levy, 2021](#)), electoral outcomes ([Fujiwara et al., 2024](#)), electoral accountability ([Enríquez et al., 2024](#)), and collective action

(Enikolopov et al., 2020; Qin et al., 2024).⁹ We examine how the intervention and rhetoric of politicians change the language used by other participants on social media.

Second, we contribute to the literature on how online platforms shape political discourse. Most of the literature has emphasized how demand-side factors such as platform algorithmic features and users' behavioral or cognitive biases drive partisanship and polarization on social media.¹⁰ For instance, contact with users sharing similar beliefs and worldviews reinforces extreme views (Bakshy et al., 2015; González-Bailón et al., 2023; Halberstam and Knight, 2016; Levy, 2021; Nyhan et al., 2023); anonymity disinhibits users (Ederer et al., 2025); and controversial and hateful speech generates higher engagement (Beknazar-Yuzbashev et al., 2025; Giavazzi et al., 2024; Guess et al., 2023).¹¹ We focus instead on how social media platforms enable prominent actors to communicate directly with large audiences and examine how such communication shapes expressions of partisanship and polarization and may thereby accentuate political and social divisions.

Third, we broadly contribute to understanding how politicians' rhetoric alters public discourse and policy. Closest to our paper is work on what politicians choose to say and with what consequences. For instance, politicians can embolden the expression of views that voters had previously suppressed (Bursztyn et al., 2020) and shift the behavior of public employees (Grosjean et al., 2022; Spenkuch et al., 2023). We contribute to this literature by showing directly how politicians' rhetoric reshapes the language used by other participants in online debates. This mechanism also speaks to the broader literature on the rise and consequences of divisive and populist politics (Funke et al., 2023; Guriev and Papaioannou, 2022).

The remainder of the paper is as follows. Section 2 describes our data. In Section 3, we document patterns and trends (or the absence thereof) in the timing of political interventions. Section 4 motivates and describes our empirical strategy. In Section 5, we present and discuss our main results. In Section 6, we investigate the mechanisms through which political interventions polarize public debates and provide evidence on the role of the

⁹A broad literature shows how traditional media affect political participation (Angelucci et al., 2024; Lenz and Lawson, 2011; Wang, 2025), political preferences (Couttenier et al., 2024; Wang, 2021b; Zhuravskaya et al., 2020), electoral outcomes (Adena et al., 2015; Amarasinghe and Raschky, 2022; Ash et al., 2024; DellaVigna and Kaplan, 2007; Martin and Yurukoglu, 2017; Wang, 2021a), as well as ethnic hostility and violent behavior (Adena et al., 2015; DellaVigna et al., 2014; Yanagizawa-Drott, 2014).

¹⁰Complementary literatures study how social media reshape news production (Cagé et al., 2022; Hatte et al., 2022) and focus on the issue of misinformation (e.g. Acemoglu et al. (2010); Angelucci and Prat (2024); Guess et al. (2020); Guriev et al. (2023); Henry et al. (2022); Jackson et al. (2022); Mattozzi et al. (2023)).

¹¹Recent experimental studies on Facebook and Instagram document substantial levels of partisan segregation online (González-Bailón et al., 2023) but find no link with polarization in the short run (Nyhan et al., 2023). However, the null result on polarization has come under recent scrutiny when it was revealed that Facebook may have manipulated the algorithm during the experiment (Thorp and Vinson, 2024).

supply of partisan rhetoric in shaping the public debate. We conclude in Section 7.

2 Data

We now discuss the construction of our database of online public debates. Supplementary Appendix B includes all the details.

2.1 Mass shooting events

Mass shooting events have been shown to affect polarization among both voters and lawmakers (Barilari, 2026; Yousaf, 2021). We aim to uncover whether political communication on social media contributes to this polarization. We focus on the 57 mass shooting events that appear on the front page of *The New York Times*¹² between March 1, 2016 and October 27, 2022.¹³ From a measurement point of view, this set of highly salient mass shooting events offer several advantages. They provide a homogeneous set of events that: (i) generate a high volume of tweets, so that we can measure variation in language; (ii) have a well-defined and unanticipated starting time, so that we can identify the first event-related tweet (by any user) and the first event-related tweet by a politician; and (iii) are not directly determined (in their timing, location, and specific characteristics) by the actions or decisions of politicians.

For each event, we retrieve: (i) the *start time* (the start of the shooting episode); (ii) the *end time* (the time at which the shooter is either arrested or dead: on average, 150 minutes after the start time, with a median of 10 minutes, for the 52 events with a recorded end time); and (iii) other characteristics, such as location (a school in 21% of the events; a business in 19%), the race of the shooter (white in 51% of the events), and the number of victims (on average, 7.7). Appendix Table B.1 lists the 57 mass shooting events in our sample, including location, timing, location type, shooter ethnicity, and the number of deaths.

2.2 Tweets

We search for all event-related tweets following a methodology similar to Demszky et al. (2019). For each event, we define two sets of keywords: one related to the circumstances of the event and the other to its geographic location. A tweet is related to that event if it contains at least one word from each set of keywords.

¹²We extract this information from *The New York Times* archive <https://archive.nytimes.com/>

¹³For the time period 2021–2022, we check the overlap between events on the front page of *The New York Times* and events on the front page of other popular publications (*The Washington Post*, *The Los Angeles Times*, *The Wall Street Journal*, and *USA Today*).



For each tweet related to an event, we retrieve: the exact time of posting, the text of the tweet, the tweet ID, and information about the posting account, including: username, user handle, profile description, and user ID. Figure 1 shows an example of the raw data retrieved from a tweet and posting account’s profile.

For each event, we identify the first related tweet and all related tweets posted within the following seven days. We refer to this sequence of tweets as the event’s *public debate*, and we define the *onset* of the public debate as the time of the first tweet. We exclude retweets and non-English language tweets. We retain a total of 4.75 million tweets (83,308 tweets per debate, on average).

2.4 Measures of polarization of the public debate

7

or “bag of phrases,” $\mathcal{B}_i$, where each word or phrase¹⁴ in the tweet is an element of the set $\mathcal{B}_i$. For each outcome y denoting partisanship or issue polarization, the corresponding dictionary $\mathcal{D}_y$ contains a set of words or phrases. We define y_i as equal to 100 if $\mathcal{B}_i \cap \mathcal{D}_y \neq \emptyset$ and 0 otherwise.

Partisanship. We measure partisanship by whether the tweet contains phrases from a dictionary of 78 partisan phrases identified by [Gentzkow et al. \(2019\)](#) (the complete list is in Supplementary Appendix B.2.1). This list includes the top 10 two-word most Republican and top 10 two-word most Democratic phrases for each Congressional Session from 2005 through 2016. It is compiled with methods aimed at selecting phrases that most enable “an observer to infer a congressperson’s party from a single utterance” ([Gentzkow et al., 2019](#)). In the full dataset, 2.8% of all tweets are partisan.

Issue polarization. We measure issue polarization by whether the tweet contains words from the dictionary of 83 words in [Simchon et al. \(2022\)](#)’s issue polarization dictionary. This dictionary is devised to isolate words that capture an “ideological, policy-based political divide,” in contrast with “affective polarization” (see Section 5.5). In the full dataset, 19.9% of all tweets are issue-polarized.

Model-based measures. To assess the robustness of our results, we construct a model-based measure of issue polarization by using LLM-generated annotations to fine-tune a DeBERTa classifier. This measure is designed to capture more semantically varied forms of issue-polarized discourse that may not contain dictionary terms. We use this measure as a complementary outcome to assess whether the results are robust to a broader, semantic definition of issue polarization. In the full dataset, 26.2% of tweets are classified as issue-polarized by this measure. The model-based measures also allow us to distinguish left- and right-aligned issue-polarized discourse and to classify tweets according to the specific issue domains being discussed. We consider nine issue domains, including gun regulation, crime and policing, mental health, immigration, or terrorism (see full list in Supplementary Appendix B.3) and a “no domain” category.

We evaluate the model-based measures against human annotations using two samples, one annotated for binary issue polarization and ideological direction, and the other annotated classified into the ten-category issue domains. For issue polarization, GPT-4.1-mini, the fine-tuned DeBERTa classifier, and three additional open-weight LLMs achieve Cohen’s κ values between 0.75 and 0.80 relative to the human annotations. GPT-4.1-mini and DeBERTa also achieve κ values of approximately 0.80 for the ten-category issue-

¹⁴As we detail below, some of our dictionaries are made of single words. Others are made of longer phrases, containing 2 to 5 words.

domain classification. These results indicate that the model-based measure corresponds closely with the human annotations and is not specific to GPT-4.1-mini. Supplementary Appendix B.3 provides full details on the annotation procedure, classifier construction, and validation.

We focus primarily on the dictionary-based measures in our main analysis. Dictionary-based measures are transparent and readily replicable, particularly when the construct is clearly defined, such as partisanship and issue polarization, and the dictionary combines data-driven methods with domain expertise and is validated across datasets (Hasan et al., 2025). They also provide tweet-level measures that are stable to variations in the underlying samples and corpora, and have a straightforward interpretation as the probability that any given tweet contains words from a given dictionary. Ash and Hansen (2023) further argue that dictionary-based measures may reduce the risk of correlated measurement error in downstream econometric analyses because they are less likely to incorporate extraneous features of the text and implicitly regularize the measurement process. While we focus on these more conservative measures, we show that our results are robust to using the broader model-based measure of issue polarization.

2.5 Interventions

Political interventions. To identify when politicians *intervene* in a public debate, we compile a list of all U.S. politicians with more than one million Twitter followers as of July 11, 2022.¹⁵ These accounts are selected from a list of all past and present U.S. presidents and members of Congress, presidential candidates, state governors, and mayors of the top 50 cities by population, between 2016 and 2022. This search returns 70 Twitter accounts (listed in Supplementary Appendix B, Table B.6).

For each public debate, we identify the first and subsequent tweets originating from the accounts of these politicians. We call these tweets the *political interventions* in the public debate. 52 out of 57 public debates feature a political intervention. The 52 first interventions originate from 29 different accounts. In 73% of cases, the first political intervention is by a Democrat. The politicians who most frequently intervene first in a debate are Rep. (D-CA) Ted W. Lieu (6 first interventions), Rep. (D-CA) Nancy Pelosi (4), and Rep. (D-CA) Eric Swalwell (4).

Most debates have interventions from multiple politicians. The median number of political interventions is 7, and the mean is 10.¹⁶ The politicians who intervene most

¹⁵We consider the robustness of our results to using alternative cutoffs of 500,000 or 200,000 followers.

¹⁶On average, a debate has 7.6 (median: 5) interventions by Democratic politicians and 2.5 (median: 1) by Republicans.

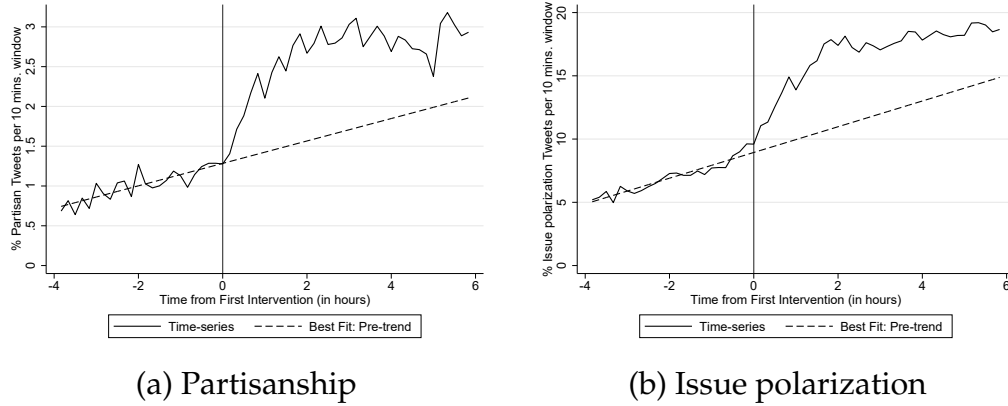


Figure 2: First political intervention, partisanship and issue polarization: raw time series

These figures plot the percentage of tweets containing partisan (panel (a)) or issue-polarized content (panel (b)) averaged over 10 minutes intervals, for events that receive a political intervention. The vertical line indicates the first political intervention in the debate.

overall are President Donald Trump (42 total interventions), the New York City Mayor¹⁷ (41), and Sen. (D-NY) Chuck Schumer (39).

Figure 2 illustrates with raw data the core message of the paper: the partisan and issue-polarized content of the public debate disproportionately increases after the first political intervention, relative to a simple time trend fitted to the pre-intervention period. A natural concern is that the timing of the political intervention may systematically coincide with specific junctures in the event itself, or be strategically timed with changes in the debate. We discuss in greater detail the patterns and timing of political interventions in Section 3 and implement two-way fixed effects and event-study methodologies (described in Section 4) in Sections 5 and 6.

Interventions by celebrity influencers. We construct a comparable list of prominent influencers who are not politicians. Thousands of athletes, artists, and business celebrities have over 1 million followers, with some playing a very active role in public debates. We retrieve a list of the top 100 accounts by followers (all with more than 25 million followers) and eliminate the accounts of politicians (including non-U.S. and non-active politicians), associations and organizations.¹⁸ This leaves us with 61 individuals. The final list (Supplementary Appendix B.4, Table B.7) includes celebrities from a variety of backgrounds, such as Elon Musk, Justin Bieber, Taylor Swift, Ellen De Generes, LeBron James, or Bill

¹⁷This account appeared under the name Bill de Blasio (D) until December 31, 2021 and Eric Adams (D) from January 1, 2022.

¹⁸For the potential role of the NRA, see footnote 21.

3 How and when politicians intervene in the public debate

In this section, we describe the patterns of political interventions in the public debate. We first focus on the timing of political interventions and then turn to a brief description of their content.

3.1 Timing of interventions

Three features of the timing of political interventions emerge from the data. First, the timing of the first intervention varies substantially across events. Second, political interventions exhibit no systematic sequencing relative to interventions by other politicians, prominent influencers, or traditional news media organizations. Third, the timing of political interventions is not systematically associated with observable characteristics of the event, or of the public debate, including its partisan and issue-polarized content and tweet volume. Together, these patterns inform the empirical strategy described in Section 4.

First political intervention. As noted above, 29 different politicians are the first to intervene across the debates. The timing of these first interventions also varies substantially. The median time of the first political intervention (the vertical solid line in Figure A.1) is 135 minutes after the onset of the debate; the mean (vertical dashed line in Figure A.1) is 226 minutes. Both are well after the onset and the peak of the debate (around 80 minutes after the onset; see Section 2).

Appendix Figure A.2a reveals great variability in the timing of the first political intervention with respect to key event and debate timelines. The earliest (first) intervention occurs 42 minutes after the start of the event (30 minutes after the onset of the debate); the latest as much as 1431 minutes (almost 24 hours) after the start of the event. The 25th percentile of the distribution of the timing of the first intervention is at 106 minutes after the start of the event and the 75th percentile is more than 4.4 hours after. The first political intervention occurs before the end of the event (arrest, incapacitation, or death of the shooter) in only a handful of cases, with the median intervention occurring 114 minutes after the end of the event. The figure also shows that *all* political interventions occur

¹⁹None of these accounts belongs to political commentators within traditional news media, who have much fewer followers, or to personalities who could be described as political influencers, i.e., political commentators who rose to notoriety mostly through their online commentary (e.g., Charlie Kirk and Scott Presler). These political influencers were relatively marginal at the time of our study, only reaching 1 million followers on Twitter at the end of our sampling period (March 2016–October 2022). They are therefore not considered in our analysis.

after the first intervention by a news media organizations²⁰ (90 minutes later, at the median), but politicians first intervene in the debate well before other influencers: there is a 235 minutes median delay between the first tweet by a politician and the first tweet by another influencer, with a very wide dispersion (interquartile range of 572 minutes).

Subsequent political interventions. Most events receive multiple political interventions, averaging 10 interventions per event. Four debates receive a single political intervention, and 42% have 4 or fewer political interventions (see Figure A.3 in Appendix). Only 18% of the events receive 20 or more political interventions, with the maximum (35) reached for the Pittsburgh synagogue shooting in 2018. The multiplicity of political interventions raises the question of whether politicians strategically react to one another.

If politicians strategically monitored and responded to one other, we would expect subsequent interventions to immediately follow the first one, and to display much reduced variability in their timing compared with the first intervention. Figure A.2b shows that this is not the case. The interquartile range of the second, third, . . . , and up to sixth intervention is comparable to that of the first one, and is even larger for subsequent interventions. There are also substantial delays between interventions, with the median second intervention occurring 205 minutes after the onset of the debate (i.e., as much as 70 minutes after the median first intervention).

Event and debate characteristics and political interventions. Table 1 shows how the timing of the first political intervention correlates with a broad set of characteristics of: (i) the intervening politician, including: gender, party affiliation, follower count, and the partisan and issue-polarized content of their tweet (Panel A); (ii) the event itself, including: race of the shooter, location of the shooting, and number of casualties (Panel B); and (iii) the public debate, including: volume of tweets, likes, retweets, and indicators of follower count of the twitter users engaged in the debate (Panel C). Columns 1 and 2 display the mean and standard deviation of the timing of the first political intervention for different values of each characteristic. Most important for our analysis, the results show that the timing of the intervention is unrelated to the partisan or issue-polarized content of the political intervention itself. Out of the 20 characteristics of the politician, event, and debate, only one—the party affiliation of the politician—is associated with a statistically significant difference in the timing of the intervention (as shown in Column 3).²¹

²⁰We use an approach similar to the approach used to select politicians’ and celebrities’ accounts to identify 76 accounts of the most influential English-language news organizations covering U.S. events on Twitter. Appendix B.4 Table B.8 contains the full list of media organizations accounts. This list includes, e.g., @cnbrk, @cnn, @nytimes, @FoxNews, and @Reuters.

²¹We also rule out the possibility that political interventions arrive systematically in response to tweets by the NRA (National Rifle Association), which could themselves spur more polarized language. The NRA

Table 1: Balance test on timing of political intervention

Dependent Variable (DV): Intervention Time			
Independent Variable (X)	Mean of DV when:		Difference
	X=0	X=1	
Panel A: Politician characteristics			
Male	198.66 (152.35)	239.49 (276.25)	40.82 (59.49)
Republican	266.79 (264.60)	104.19 (74.63)	-162.61*** (47.23)
SUPPLY: Partisanship	221.89 (255.97)	244.01 (180.97)	22.13 (68.19)
SUPPLY: Issue polarization	204.45 (242.06)	291.22 (238.94)	86.77 (75.75)
SUPPLY: Tribalism	251.15 (301.21)	194.61 (135.47)	-56.53 (62.73)
SUPPLY: Affective polarization	210.68 (301.25)	240.46 (174.92)	29.78 (68.99)
Followers Count: High	240.91 (273.72)	192.92 (149.95)	-47.98 (58.94)
Panel B: Event characteristics			
Shooter race: White	222.52 (204.52)	210.50 (269.3)	-12.03 (67.02)
Shooter race: Black	206.35 (247.19)	249.24 (223.76)	42.88 (76.81)
Shooter race: Other	221.42 (254.98)	195.81 (190.27)	-25.61 (69.34)
Shooting location: School	209.55 (172.52)	239.39 (381.74)	29.84 (111.17)
Shooting location: Business	227.24 (254.33)	166.75 (88.64)	-60.49 (48.71)
Shooting location: Community	226.68 (251.72)	162.20 (86.89)	-64.48 (48.55)
Shooting deaths: High	237.93 (277.39)	201.88 (148.77)	-36.05 (59.12)
Shooting length: Long	187.18 (232.66)	298.46 (228.81)	111.28 (85.52)
Panel C: Twitter characteristics before intervention			
Tweets Volume: High	223.00 (266.43)	233.89 (174.94)	10.89 (62.62)
Poster’s Followers: High	228.01 (282.26)	223.15 (164.63)	-4.86 (62.03)
Poster’s Followings: High	233.41 (279.38)	213.51 (163.75)	-19.90 (61.45)
Event’s Like Count: High	217.60 (267.64)	245.35 (176.92)	27.75 (62.6)
Event’s Retweet Count: High	228.39 (285.16)	222.82 (165.11)	-5.57 (62.66)

The unit of observation is an event. Column 1 (respectively, 2) shows the mean and standard deviation of the time (in minutes from the onset of the debate) of the first political intervention when the variable in the corresponding row is equal to 0 (respectively, 1). Column 3 displays the coefficients estimated from separate OLS regressions (with robust standard errors) of the timing of the first political intervention (dependent variable) on each row variable (independent variable).

Table A.1 in Appendix conducts a similar analysis to explore whether the *number* of political interventions depends on the politician, event, and debate characteristics. The number of interventions covaries significantly with only four out of 20 characteristics. There are substantially more interventions in the events with higher victim count, when the shooter is not Black, and when the first politician to intervene is a male or Republican. However, the partisan and issue-polarized content of the first political intervention, the online popularity of the debate, and the follower count of the first politician to intervene are uncorrelated with the number of subsequent interventions. In our empirical investigation, event fixed effects will account for any heterogeneity in fixed politician-, event-, or debate-level characteristics that are associated with differences in the timing or number of political interventions.

Overall, the evidence so far suggests no systematic timing of political interventions relative to one another, to key points in the event timeline, or to debate characteristics. In Section 5, we use an event-study methodology to further show that political interventions are also unrelated to immediate pre-trend *changes* in the characteristics of the online debate. We first briefly describe the content of political interventions.

3.2 Rhetorical content of political interventions.

We characterize how political interventions differ rhetorically from interventions by celebrities, news media organizations, and regular users. The key finding is that political interventions “supply” partisan worldviews. Column 1 of Table A.2 shows that the first political intervention by a politician is 18 to 19 percentage points more likely to contain partisan language compared to a tweet by a regular user prior to a political intervention (the excluded category) or to the first intervention by a celebrity or a traditional news media organization. These differences are all statistically significant at the 1% level. Results in Column 5, which consider all (not just first) interventions by politicians, celebrities, and news media, show similar results both in terms of magnitude and statistical significance. Columns 2 and 6 show that interventions by politicians are also more likely to contain issue-polarized language compared to pre-treatment tweets by regular users (by 16 percentage point when considering the first intervention or 24 percentage points when considering all interventions). However, the difference in issue-polarized content between politicians and celebrities or news media is insignificant for the first intervention (as indicated at the bottom of the table).

Further results in Table A.3 confirm that partisan rhetoric is the hallmark of politicians’

has less than one million followers on Twitter and is known to avoid responding to mass shooting events (Merod, 2016). In fact, in our dataset, the NRA intervenes only 18 times across 10 debates (out of 57).

interventions. While political interventions also contain more tribal and affectively polarized language²² compared with tweets by regular users, these rhetorical patterns do not distinguish politicians from celebrities or news media in a statistically significant way. As for hate speech, first tweets by politicians are statistically undistinguishable from those of regular users.

We can also dig within the dictionaries to further explain what characterize the partisanship and issue-polarization of political interventions. Table A.4 reports the 15 dictionary keywords most frequently used by regular users and compares their prevalence across account types. Among these common terms, political interventions are particularly characterized by a significantly greater use of the words “law”, “dem”, “republican”, “gun violenc”, and “senat”.

4 Empirical strategy

The discussion and results in Section 3 underpin our strategy. To fix ideas, one way to think of the process generating an intervention by a specific politician is as one that includes deterministic, stochastic, and idiosyncratic components. The start of a public debate triggers a stochastic process that informs politicians of the event. While all politicians eventually become informed, the exact time at which a politician becomes informed has a stochastic component that depends on their individual and event-specific characteristics, but also on unrelated variables such as their schedule of travels and business on the day of the event. From the moment the politician becomes informed, they may, depending on politician-event-specific characteristics, begin the process of formulating their intervention. Once the politician has formulated an intervention, then a stochastic process pins down the exact time when they or their staff are able to enact the intervention (post the tweet). The exact delay between becoming informed and tweeting is therefore also in part stochastic and varies across politicians and events.²³ The stochastic nature of these delays is what generates the quasi-random nature of the timing of political interventions we documented in Section 3 and allows us to study the effect of such interventions on the public debate.

We conduct our analysis at the tweet-event level. We normalize time within each public debate so that the onset of the public debate is time 0. Therefore, each tweet in a debate is posted at a time t from 0 to T (7 days later). Our objective is to estimate whether

²²See Section 5.5 for details on these measures.

²³Even before socials, readers may recall the process that led President G.W. Bush to first become informed, processing the information, and later choosing how to intervene in the immediate aftermath of the terrorist attacks on September 11, 2001.

a political intervention in a public debate changes the probabilities that subsequent tweets in the same debate contain partisan or issue-polarized language. To do so, we begin by estimating the effect of the *first* political intervention.

For each tweet i posted at time $t \in [0, T]$ in event e 's debate, we define $D_{i,e,t}^{(u,v)}$ as a dummy variable equal to 1 if and only if $t \in (u, v)$ (i.e., if the tweet is posted between times u and v). Let p_e be the time of the first political intervention in event e 's debate.

4.1 Two-way fixed effect specifications

Our baseline two-way fixed effect analysis estimates:

$$y_{i,e,t} = \beta D_{i,e,t}^{(p_e, T)} + \eta_e + \theta_t + \zeta t + v_{i,e,t} \quad (1)$$

where $y_{i,e,t}$ measures the semantic content of tweet i , with expressions of partisanship and issue polarization as our primary outcomes, η_e is an event-specific fixed effect, θ_t denotes a set of 30-minutes intervals fixed effects, and ζ captures the influence of time trends, measured in continuous time since the onset of the event. In this specification, the pre-intervention period $(0, p_e)$ is the excluded reference window. We address inference issues with the estimation of two-way fixed effect models (Borusyak et al., 2024; De Chaisemartin and d'Haultfoeuille, 2020; Sun and Abraham, 2021) in Section 5.4.

The nature of the debate, including in terms of partisanship and issue polarization, and the characteristics of the political intervention may vary across events, for example depending on victim count, or location. Although, as we discussed, the timing of the intervention does not systematically vary with event characteristics, we still include event-specific fixed effects, η_e . These fixed effects account for any unobserved heterogeneity in event-specific debate or political interventions. The dispersion of tweet content may also vary greatly across events. For this reason, we cluster standard errors by event.

Another immediate concern is that the nature of the public debate naturally changes over time. These changes could stem both from a change in the nature of the debate (for example, tweets may organically become more or less partisan as users are exposed to more arguments) or from a compositional change among Twitter users, as different people join or exit the debate as time goes by. For example, if more partisan users become interested in the event as the debate gains traction, the debate may naturally become more partisan over time. We account for these potential changes in the debate using intervals of time-since-the-onset-of-the-debate fixed effects. In (1), θ_t denotes a set of 30-minutes intervals fixed effects from the start of the event. We consider alternative intervals of 15 or 60 minutes in robustness analyses. We also allow a linear time trend to account for

continuous changes in the debate over time.

Because the volume of tweets varies over time, the dispersion of the error term may also vary over time, which may affect the precision of our estimates. We therefore cluster standard errors over the time intervals dimension, in addition to the event dimension. These standard errors adjustments correct inference issues related to the serial correlation of the error term $v_{i,e,t}$ within an event over time, and within intervals of time since debate onset, across events.

In further robustness checks, we account for the fact that users posting immediately after a political intervention may not have had time to read the intervention, or may have formulated their tweet before the intervention and experienced a small delay in posting it, so that the tweet appears in our dataset *after* the intervention but could not have been affected by it. To account for this mis-classification risk, we include a separate dummy $D_{i,e,t}(p_e, p_e + \epsilon)$ for tweets posted immediately after the intervention and show that our results are unaffected when including this dummy for either ϵ equal 1 or 2 minutes.

A potential limitation of specification (1) is that, as noted in Section 2, political interventions arrive at different times between time 0 and T for different events. This implies that the pre-intervention period, $(0, p_e)$, and the post-intervention period, (p_e, T) , are heterogeneous across events, so that β averages across different time periods, which may complicate the interpretation of the results.

To harmonize comparison windows, we estimate, for $a, b, c > 0$,

$$y_{i,e,t} = \alpha D_{i,e,t}^{(0,p_e-b)} + \beta_1 D_{i,e,t}^{(p_e,p_e+a)} + \beta_2 D_{i,e,t}^{(p_e+a,p_e+c)} + \beta_3 D_{i,e,t}^{(p_e+c,T)} + \eta_e + \theta_t + \zeta t + v_{i,e,t} \quad (2)$$

$D_{i,e,t}^{(0,p_e-b)}$ captures all tweets from the onset of the debate until b minutes *before* the intervention; $D_{i,e,t}^{(p_e,p_e+a)}$ captures all tweets in the a minutes *after* the intervention; $D_{i,e,t}^{(p_e+a,p_e+c)}$ captures all tweets between a and c minutes after the intervention; and $D_{i,e,t}^{(p_e+c,T)}$ captures all tweets posted *later* than c minutes after the intervention. The coefficients of interest, β_1 and β_2 , capture the changes in the partisan or issue-polarized content of tweets posted in the a minutes after the intervention (short-term), and between minutes a and c after the intervention (medium-term), compared to the comparison window of length b before the intervention. The coefficient β_3 , which captures this change later than c minutes after the intervention until T (long-term) is difficult to interpret because, as time goes on in the debate, further interventions within the debate and other news events may influence the rhetoric of the public debate. Based on the event-studies analysis of the dynamics of the effect, we set a, b and c to 60 minutes.

In the classical potential outcome framework, the assumption required to causally

identify the effect of a political intervention is that, in the absence of a political intervention at time p_e , after controlling for the time elapsed since the onset of the debate, as well as both time intervals since the onset of the debate and event fixed effects, the partisan and issue-polarized content of the public debate would remain unchanged (Equation (1)), or at least would remain unchanged between b minutes before the intervention and $a + c$ minutes after (Equation (2)). This assumption would be violated if politicians systematically timed their interventions in response to the changes in public debate we hope to characterize. We address potential violations of the identification assumption in several ways. First, the high-dimensional time fixed effects ensure that we are comparing tweets posted after a political intervention (treated tweets) to tweets posted in a debate about a similar event but in which a politician has not intervened (yet, or at all) within the same (short) interval of time after the onset of the debate. The inclusion of event-specific time trends in our robustness analysis further enables us to account for different linear dynamics in the nature of tweets across debates. Second, we have already discussed how the timing of an intervention is uncorrelated with any of the debate characteristics, including its prominence (volume of tweets, number of likes and retweets) or the profile of users engaged in it. Moreover, we begin our analysis with an event-study analysis that shows that there are no pre-trend differences in the partisan or issue-polarized content of the debate, as well as in the volume of public debates, before a political intervention.

4.2 Event-Study specification

An event-study analysis offers several advantages over the two-way fixed effect approach described thus far. First, it allows us to test for pre-trends in outcomes before a political intervention. Second, it enables us to study how political interventions affect the public debate over time, rather than averaging over the whole window from the time of the intervention to either the whole post-intervention period or a , or $a + c$ minutes later, and to estimate precisely when the effect of a political intervention materializes (thereby also motivating our choice of parameters a and c). To smooth out noise in daily observations, we estimate parameters for 10-minutes bins, starting from 180 minutes before the political intervention and ending 300 minutes after it. Therefore, the event-study specification estimates (where time is expressed in minutes):

$$y_{i,e,t} = \alpha D_{i,e,t}^{(0,p_e-180)} + \sum_{\tau=-18}^{-2} \beta_{\tau} D_{i,e,t}^{(p_e+10(\tau),p_e+10(\tau+1))} + \sum_{\tau=0}^{29} \beta_{\tau} D_{i,e,t}^{(p_e+10\tau,p_e+10(\tau+1))} \quad (3)$$

$$+ \beta_T D_{i,e,t}^{(p_e+300,T)} + \eta_e + \theta_t + \zeta t + v_{i,e,t}.$$

As in (2), we include fixed effects for each 30-minutes interval from the onset of the debate, event fixed effects, and a linear time trend. The omitted time bin includes all tweets posted in the 10 minutes prior to the political intervention (i.e., in the interval $(p_e - 10, p_e)$).²⁴

4.3 Marginal effects of political interventions

So far, we focused on estimating the effect of the first political intervention in the public debate. Most public debates receive multiple interventions. Therefore, we estimate the following equation to capture the marginal effect of subsequent political interventions:

$$y_{i,e,t} = \gamma f\left(\sum_0^t D_{i,e,t}^{(0,t)}\right) + \eta_e + \theta_t + \zeta t + v_{i,e,t}. \quad (4)$$

For each tweet i posted at time t , the treatment of interest is the sum of political interventions that have occurred until t .²⁵ We characterize f by including higher order polynomials of the number of cumulative interventions and plotting the marginal effect of each intervention.

5 Political interventions polarize public debates

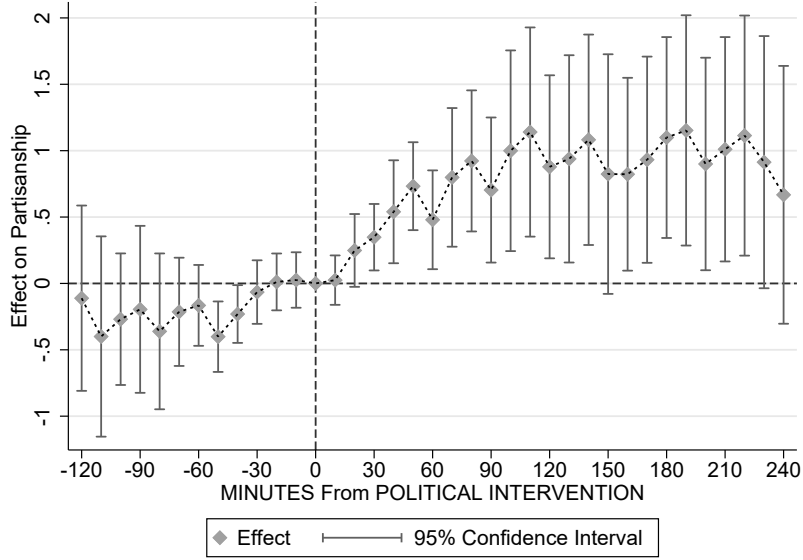
We now show that political interventions increase the partisan and issue-polarized content of the public debate. The first political intervention has the largest effect, but subsequent interventions shape the debate further, albeit at a decreasing rate. We also discuss the robustness of our results to alternative specifications, as well as additional results.

5.1 Event-study results

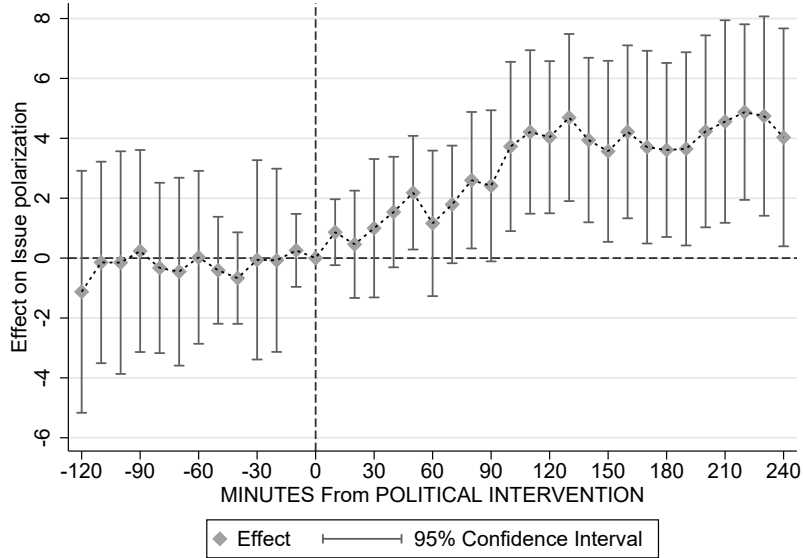
Figure 3 displays the estimates of β_τ , $\tau \in \{-12, \dots, 24\}$ in (3) for partisan (Panel (a)) and issue-polarized (Panel (b)) language. The results show that the timing of the first political intervention is unrelated to prior trends in the partisan or issue-polarized content of the public debate. Yet, the probability that a tweet in the public debate contains partisan or issue-polarized language significantly increases after the first political intervention.

²⁴Borusyak et al. (2024) show that in some settings with staggered treatment where each unit is treated only once, econometric models with unit and time fixed effects are unable to identify a unit-specific linear trend. This happens in fully dynamic settings where all units are eventually treated. Our framework is different because our estimation sample includes events that are never treated, in the sense that no politician posts a tweet related to these events.

²⁵For instance, suppose that a public debate receives two political interventions, one 60 minutes into the debate, and the other 90 minutes into the debate. Then, the variable $\sum_0^t D_{i,e,t}^{(0,t)}$ takes value 0 for $t < 60$, 1 between $t = 60$ and $t = 90$, and 2 for $t > 90$.



(a) Partisanship



(b) Issue polarization

Figure 3: Event-study results

These figures plot OLS coefficients with 95% confidence intervals. The plotted coefficients are the β_τ coefficients associated with each 10 minute window, described in Equation (3). The outcome variable is Partisanship in panel (a) and Issue Polarization in panel (b). Standard errors are two-way clustered at the event level and at the level of 30 minutes time intervals since the onset of the debate.

The effect of the first political intervention on partisanship is immediate. The probability that a tweet contains partisan language is .25 percentage points higher in the 10–20 minutes window after the intervention, compared to before. The effect builds up, in-

creasing to .54, .80, and 1.00 percentage points in the 30–40, 60–70, and 90–100 minutes windows after the intervention. After 90 minutes, the effect remains stable, between .8 and 1.1 percentage points, and persists at this level in the following three and a half hours (up to 210 minutes).

In comparison, the effect of the political intervention on issue polarization takes more time to materialize but is persistent after that. The probability that a tweet contains issue-polarized language is 2.18 percentage points higher in the 40-50 minutes window after the intervention, compared to before, an effect that is statistically significant at the 5% level. The effect increases steadily, to 3.72, 4.69, and 4.87 percentage points in, respectively, the 90-100, 120-130, and 210-220 minutes windows after the intervention and persists at this level in the following hours.

5.2 Two-way fixed effects results

Main Results. Column 1 in the first panel of Table 2 displays the estimate of β from Equation (1) for partisanship content of a tweet in the 7 days after the first political intervention in the debate. The probability that a tweet contains partisan language increases by .71 percentage points (significant at the 1% level), which represents an approximately 65% increase relative to the pre-intervention mean.

The second panel of Table 2 displays the estimates associated with Equation (2). The probability that a tweet contains partisan language increases by 0.47 percentage points (a 43% increase relative to the pre-intervention mean) in the first 60 minutes after the first political intervention, compared to the 60 minutes before the intervention, and by .96 percentage points (an 88% increase relative to the pre-intervention mean) in the 60-120 minutes after the intervention. Both coefficients are statistically significant at the 1% level. The effect of a political intervention on partisanship is long-lasting: partisanship is 0.98 percentage points higher more than two hours after the political intervention relative to 60 minutes before.²⁶

Column 2 displays the corresponding estimates for issue polarization. The probability that a tweet contains issue-polarized language increases by 2.43 percentage points after the first political intervention, which is statistically significant at the 1% level and amounts to a 31% increase relative to the pre-intervention mean (first panel). The second panel confirms that, although the effect of political interventions on issue polarization is slightly slower to materialize compared with partisanship, the probability that a tweet contains issue-polarized language increases by 1.31 percentage points (a 17% increase

²⁶The coefficients estimated from (1) and (2) are not directly comparable to one another because the excluded comparison windows are different.

Table 2: First political intervention, partisanship, and issue polarization

VARIABLES	(1) Partisanship	(2) Issue polarization
Equation 1		
POST-Intervention (β)	0.705*** (0.200)	2.434*** (0.753)
Observations	4,747,621	4,747,621
R-squared	0.014	0.045
Time Trend	Linear	Linear
Equation 2		
POST-Intervention $< 1h$ (β_1)	0.465*** (0.120)	1.313*** (0.383)
POST-Intervention $1 - 2h$ (β_2)	0.958*** (0.303)	3.122*** (0.812)
POST-Intervention $> 2h$ (β_3)	0.983*** (0.361)	4.229*** (1.370)
Observations	4,747,621	4,747,621
R-squared	0.014	0.045
Time Trend	Linear	Linear
Mean DV	1.08	7.75

The unit of observation is a tweet. The dependent variable is equal to 100 (and zero otherwise) if the tweet contains partisan language (column 1), or issue-polarized language (column 2). Panel A shows the OLS estimates of β from Equation (1). Panel B shows the OLS estimates of β_1 , β_2 , and β_3 from Equation (2) with $a = b = c = 60$. Regressions exclude the 909 tweets by politicians. Standard errors are two-way clustered at the event level and at the level of 30 minutes time intervals since the onset of the debate. Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

relative to the pre-intervention mean) in the first 60 minutes after the first political intervention, compared to the 60 minutes before the intervention, and by 3.12 percentage points (a 40% increase relative to the pre-intervention mean) in the 60-120 minutes after the intervention. Both coefficients are statistically significant at the 1% level. The effect of a political intervention on both partisanship and issue-polarization is long-lasting: partisanship is 0.98 and issue-polarization is 4.23 percentage points higher two hours after the political intervention relative to 60 minutes before.

Overall, these results align with the dynamics uncovered in the event-study figure: an immediate, large, and persistent effect of a political intervention on partisanship, and a slower, but equally persistent effect on issue polarization in the medium term.

Model-based Issue Polarization Results. To confirm that our issue-polarization results are not an artifact of dictionary construction, we re-estimate our main specification using

the model-based measure described in Supplementary Appendix B.3, which classifies each tweet with a fine-tuned DeBERTa model trained on LLM annotations. The results in Table A.5 show that our results are quantitatively and qualitatively similar using this alternative measure. The probability that a tweet is classified as issue-polarized rises by 2.14 percentage points after the first political intervention (column 1), a 24% increase relative to the pre-intervention mean and close to the 2.43-point (31%) dictionary estimate in Table 2. The dynamics also match: the effect is statistically indistinguishable from zero in the first hour, reaches 3.02 points in the 60–120 minutes after the intervention, and remains at 3.82 points thereafter.

Because the model-based measure additionally codes the ideological direction of each tweet, we can separate left- from right-aligned issue polarization. Columns 2 and 3 of Table A.5. Both rise significantly, by 1.14 points (+19%) and 0.86 points (+37%), respectively. In other words, the increase in issue polarization is driven by both sides of the political spectrum.

Which words and topics are most triggered by political interventions? In addition to our two-way fixed-effects analysis of whole dictionaries, we conduct word-by-word analogues for each term in each dictionary. This gives a clearer picture of which elements of the dictionaries are most responding to political interventions. Appendix Table A.6 shows results for the top 10 most common words in each dictionary (before the first political intervention). The stand-out words are “law”, “gun control”, and “gun violence”. Their frequency in tweets by all users increases by 65%, 108%, and 55% after the first political intervention. Yet, the response is not confined to law or gun-related language: among the top 10 words in the two dictionaries, the frequencies of “gop”, “lib”, and “business owner” rise by 14%, 40%, and 433%, respectively, pointing to a broader polarizing effect of political interventions, beyond gun policy.

We complement this word-level analysis with the model-based classifier, which assigns every tweet both an issue-polarization score and one of ten issue domains. Two patterns emerge (Appendix Table A.7 and Table A.8). On the extensive margin, interventions shift the conversation toward gun policy: the share of tweets engaging no issue domain falls by 3.71 points, while the share about gun law rises by 3.85 points (significant at the 1% level); no other domain moves significantly. Yet, on the intensive margin, the picture is different. Conditional on the domain being discussed, gun-law tweets become *less* issue-polarized after an intervention, whereas discussions of terrorism, the media, and race become *more* polarized. Together with our results characterizing political interventions (Section 3.2), these findings suggest that the partisan content of political interventions may translate into a broader politicization of how other users discuss mass-

Table 3: Number of political interventions, partisanship, and issue polarization

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)
		Partisanship			Issue polarization	
# Interventions	0.108*** (0.031)			0.802*** (0.184)		
# Interventions Squared	-0.004*** (0.001)			-0.014*** (0.005)		
Log(# Interventions)		0.245* (0.142)	0.055*** (0.018)		3.989*** (0.539)	0.220*** (0.058)
Observations	4,747,621	4,747,621	4,747,621	4,747,621	4,747,621	4,747,621
R-squared	0.014	0.014	0.014	0.046	0.046	0.045
Weights	None	None	Followers	None	None	Followers

The unit of observation is a tweet. The dependent variable is equal to 100 (and zero otherwise) if the tweet contains partisan language (Columns 1 to 3), or issue-polarized language (Columns 4 to 6). The table shows the OLS estimates of β from Equation (4) for different functional forms of γ . Regressions exclude the 909 tweets by politicians. Standard errors are two-way clustered at the event level and at the level of 30 minutes time intervals since the onset of the debate. Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

shooting events. We return to this idea in Section 6, where we examine how political interventions polarize public debates.

5.3 Marginal effects of political interventions

Table 3 displays the estimates of γ from Equation (4) for different specifications of f . In Columns 1 and 4, we estimate a quadratic relationship between the number of political interventions and our outcomes of interest. The estimated coefficients show that more political interventions have a larger polarizing effect on the public debate, but at a decreasing rate.

Appendix Figure A.4 plots the marginal effects of an OLS estimation of partisan or issue-polarized content of a tweet posted at time t as a quartic function of the number of political interventions that have occurred until t . The first few interventions have the strongest impact on the probability that the tweet contains partisan (Panel a) or issue-polarized (Panel b) language. The marginal effect of an additional intervention by a politician becomes statistically indistinguishable from zero after the 7th intervention for partisanship and the 19th intervention for issue polarization.

Given the non-linear relationship between debate polarization and the number of interventions, we specify f as a logarithmic function for ease of interpretation. Estimates in Columns 2 and 5 of Table 3 suggest that a 100% increase in the number of interventions (doubling the interventions) increases partisanship by 0.25 percentage points and issue-polarized content by 3.99 percentage points.

To address the possibility that the popularity of intervening politicians on social me-

dia may amplify their polarizing effects, we also estimate a weighted version of (4), in which we weigh the number of interventions by the cumulative number of followers of the politicians that have intervened until time t .²⁷ The results, shown in Column 3 for partisanship and Column 6 for issue polarization, are consistent with preceding results.

5.4 Robustness

Recent literature on staggered DID highlights potential issues with the two-way fixed effects estimator stemming from the fact that the estimated parameter is a weighted average of each treatment (in our context, each first political intervention) where the weights may be negative. We follow the recommended diagnostic by [De Chaisemartin and d’Haultfoeuille \(2020\)](#) and compute the weights associated with each treatment.²⁸ Figure A.5 shows that there is little variation in the weights and that only a handful (less than 10%) is negative. We also address possible inference issues related to the simultaneous inclusion of time trends and event-fixed effects in the event study ([Borusyak et al., 2024](#)). Appendix Figure A.6 shows that the results are unchanged when we exclude the time trend from Equation (3). Finally, we apply the method proposed by [Sun and Abraham \(2021\)](#) to estimate the treatment effect for each event individually. For partisanship, the mean and median treatment effects are 1.00 and 0.72, respectively. For tribalism, the mean and median treatment effects are 2.47 and 1.75. This shows that the median and average of individual treatment effects are very similar to the main estimates obtained using the two-way fixed effect methodology, suggesting little heterogeneity in treatment effects.

We implement sensitivity analyses to potential violations of the parallel trends assumption. The canonical concern here is that unit-specific time shocks may affect the outcome even in the absence of the treatment. In our setting, since treatment occurs at different moments in calendar time, unobserved time shocks that may systematically affect treated and untreated events differently (related, e.g., to the news of the day) are not of concern. A concern would arise in our setting if interventions always occurred at the same time (since the onset of the event), which may systematically coincide with key junctures in the debate that may systematically affect outcomes. We have discussed in Section 3 that this is not the case. We still implement the sensitivity analysis suggested by [Roth \(2022\)](#) and [Rambachan and Roth \(2023\)](#). The results, displayed in Appendix Fig-

²⁷For instance, suppose there is an event with only two interventions, with the first intervention occurring at $t = 60$ and the second at $t = 90$. Suppose the politician who intervenes first has 1.2 million followers, while the second politician has 3.4 million followers; then our main independent variable takes value 0 for $t < 60$, 1.2×10^6 for $60 < t < 90$, and 4.6×10^6 for $t > 90$.

²⁸We collapse our data into 30-minute windows for estimation and concentrate on the short- and medium-term effects.

ure A.7, show that the post-treatment violation of parallel trends should be more than 1.0 to 1.5 times the *maximum* pre-treatment violation in order to explain away our results. Given the granularity and high dimensionality of the data used to estimate the trends, based on tweets in real time, such large deviations seem unlikely.

To ensure that Equations (1) and (2) are capturing the effect of political interventions rather than artifacts generated by the specification, Figure A.8 displays the results of permutation inference tests (based on 1,000 replications) in which we randomly assign the timing of political interventions across events. We consider different restrictions in the timing of political interventions, to occur either at any time in the post period (left panel), to mirror our estimate in Equation (1)), or during the time intervals considered in Equation (2) (middle and right panels). The results consistently show that our effect sizes are well outside the range of estimated effects from these placebo treatments. We also note that the randomization inference results are not always centered around 0 because the sample consists of a higher proportion of post-treatment observations relative to the pre-treatment sample (the post-treatment sample is 93.2% of the overall sample).²⁹

We implement a series of additional robustness checks. First, we check that our results are unchanged when we correct for potential missclassification of tweets posted immediately after the first political intervention. To do so, we control separately for tweets posted within a one- or two-minute post-intervention buffer. The results are unchanged (Appendix Table A.9). Second, we allow the time trend since onset of the debate to vary across events by adding an event-specific time trend to our estimation. Results, displayed in Appendix Table A.10, are robust. Third, we show that our results are insensitive to varying the definition of the time dimension of the fixed effects by considering shorter (15 minutes) or longer (60 minutes) time intervals since the onset of the debate (Table A.11). Fourth, we show in Appendix Table A.12 that our results are robust when we change the set of politicians considered in the analysis. We augment our sample to include all the (first) interventions by any politician with at least 500,000 (Columns 1 and 2) or 200,000 (Columns 3 and 4) followers. Our two-way fixed effects estimates are consistent with our baseline results, suggesting that our results are not due to selective inclusion of specific politicians in our analysis set. Last, we explore the sensitivity of our results to any event in particular. Figure A.9 in Appendix displays estimates of β , β_1 , and β_2 in a series of estimations of Equation (1) and (2) when we exclude one event at a time. The coefficients are stable and similar to our main estimates across all specifications.

²⁹This means that we over-sample from the post-treatment sample in randomized inference, thus resulting in our estimates being centered around a small positive number (0.06 for partisanship for estimates of β_1 , and 0.10 and 0.05 for partisanship and issue-polarization estimates of β_2 ; β is centered around 0).

Finally, one possible concern with our identification is that our estimates could instead capture the effect of news shock, such as important new information released by news, that are released immediately before (and perhaps cause) political interventions. Three results speak against this. First, we observe no spike in news activity just before (or surrounding) the first political intervention and event studies following Equation (3) show no discontinuity in the share of debate tweets posted by news media around the first political intervention (Appendix Figure A.10). Second, news media tweets posted in the 30 minutes before the intervention are no more likely than those posted before placebo intervention times to report deaths, information on the victims, the shooter’s identity, age, or race, the weapon, or the shooter’s motive (Appendix Figure A.11). Third, our estimates are virtually unchanged when we control for recent news media activity or drop news media tweets from the sample (e.g., $\beta_1 = 0.458$ and 0.464 vs. 0.465 for partisanship; Appendix Table A.13).

5.5 Additional Results

Volume of Tweets. We estimate our event-study specification with the volume of tweets in each 10-min window as the dependent variable. The results, displayed in Figure 4, show the absence of any pre-trend in the volume of the debate prior to a political intervention, lending further credence to the main identifying assumption that politicians do not systematically time their intervention in response to immediate changes in the debate. The results also shows that the volume of tweets does not change after a political intervention, suggesting that political interventions do not radically change the online debate beyond its partisan and issue-polarized content.

Long-term effects. The results are consistent in event studies that consider a longer time horizon of 15 hours after the intervention. These results are displayed in Appendix Figure A.12, with one-hour bins to ease visualization. Partisanship increases steadily after the first political intervention, stabilizes after 4 hours, and remains significant in the long run. Issue polarization only materializes one to two hours after the intervention, increases rapidly in the first 4 hours, increases more slowly in the 4–10 hours after the intervention, and stabilizes after that.

Political interventions and online animosity. We also explore whether political interventions generate more tribal and hateful speech. We measure tribalism by whether the tweet contains words from the dictionary of 52 words expressing loyalty and betrayal, as defined by the Moral Foundations Dictionary created by [Graham et al. \(2009\)](#) and used by, e.g., [Enke \(2020\)](#) (the complete list is in Supplementary Appendix B.2.2). Expressions of loyalty and betrayal capture “people’s emphasis on being loyal to the in-group” (e.g.,

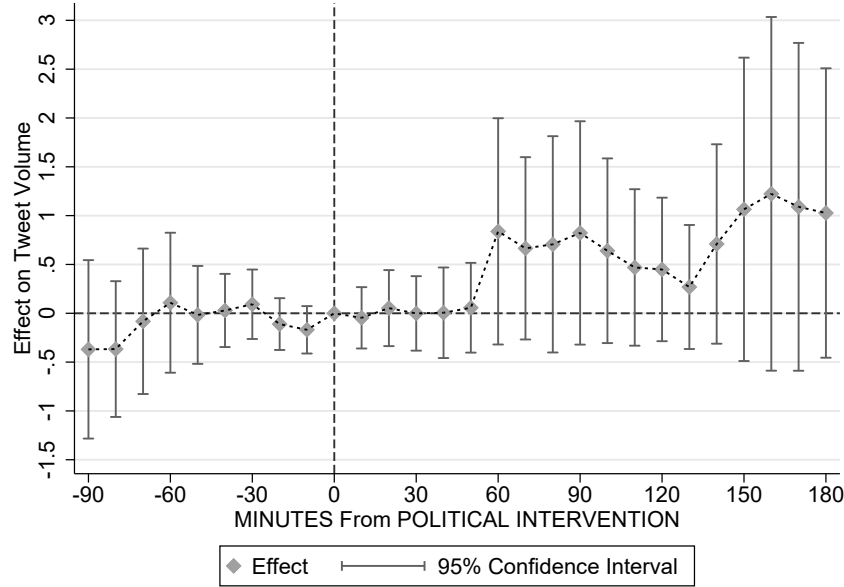


Figure 4: First Political Intervention and Volume of Tweets

The figure plots the OLS coefficients with 95% confidence intervals. The plotted coefficients are the β_τ coefficients described in Equation (3). The dependent variable is the number of tweets posted in each 10 minute window. Standard errors are two-way clustered at the event level and at the level of 30 minutes time intervals since the onset of the debate.

family or political party) (Enke, 2020) and are “related to our long history as tribal creatures”.³⁰ They are not specific to any political party and not predictive of partisan identity (Koleva et al., 2012). Expressions of loyalty and betrayal have been linked to the contemporary American “culture wars” (Enke, 2020; Graham et al., 2009; Haidt and Graham, 2007; Koleva et al., 2012).

We measure affective polarization by whether the tweet contains words from the dictionary of 86 words in Simchon et al. (2022)’s affective polarization dictionary. This dictionary is devised to isolate words that capture negative moral-emotional words. In contrast to the issue polarization dictionary (also from Simchon et al. (2022) and described in Section 2.4), this dictionary reflects the “dislike, distrust, and general animosity of political partisans”.

We measure hate speech by whether the tweet contains words from the dictionary of 3828 terms expressing hate speech and white nationalist rhetoric from Siegel et al. (2021). This dictionary contains hostile expressions directed towards an outgroup. It includes race-specific, homophobic, and misogynistic hate speech, and terms praising or associated with white nationalist groups.

We report in Appendix Table A.14 and Figure A.13 our estimations of (1)–(3) when the independent variables are the measures of tribalism, affective polarization, or hate

³⁰See <https://moralfoundations.org> (retrieved December 9, 2023).

speech. We observe a slow increase in tribal language as a result of the first political intervention, which is statistically significant in the second hour after the intervention (3.20 percentage points, 1–2h window); the average effect over the whole post-intervention period is positive but not statistically significant.³¹ However, the online debate surrounding mass shootings does not seem to become more affectively polarized or hateful after a political intervention.

We now explore mechanisms that explain why political interventions generate increases in partisanship and issue polarization, but not in affective polarization and hate speech. As we show in the next section, the key mechanism relies in the nature of the rhetoric used by politicians.

6 How political interventions polarize public debates

This section investigates the mechanisms through which political interventions shape public debates. We start by establishing the singular effect of political interventions, compared with interventions by other public figures. We then show how political interventions systematically differ in their rhetoric and show how this explains our results. Last, we show that our results are mostly due to newcomers to the debate, rather than a change in the expression of users already engaged in the debate or a change in the composition of users.

6.1 Interventions by personalities who are not politicians do *not* polarize public debates

The effects of political interventions we have documented so far may simply be due to the salience, centrality, and popularity of politicians online. Politicians are focal users with large follower counts and broad reach (Alatas et al., 2024; Bakshy et al., 2011), and their followers are more likely to consume and diffuse their messaging (Anger and Kittl, 2011). Given ideological segregation over news consumption on the internet (see, e.g., Guess et al., 2023; Levy and Razin, 2019; Levy, 2021), the effects we observe following a political intervention may not be due to the political intervention in itself, but to the news diffusion and amplification that an intervention by a focal user entails. Another possibility is that politicians are divisive by their very nature, so that their interventions naturally divide users across camps.

³¹The probability that a tweet contains tribal language is 1.59 percentage points higher in the 30-40 minutes window after the intervention, compared to before, an effect that is statistically significant at the 5% level. The effect increases steadily, to 2.61, 4.25, and 4.79 percentage points in, respectively, the 60-70, 90-100, and 120-130 minutes windows after the intervention.

Table 4: Interventions by celebrities, partisanship and issue polarization

VARIABLES	(1) Partisanship	(2) Issue polarization
Equation 1		
POST-CELEBRITY (β)	-0.144 (0.231)	0.616 (1.069)
Observations	4,748,501	4,748,501
R-squared	0.014	0.045
Time Trend	Linear	Linear
Equation 2		
POST-CELEBRITY $< 1h$ (β_1)	0.166 (0.223)	0.316 (0.732)
POST-CELEBRITY $1 - 2h$ (β_2)	-0.161 (0.242)	0.272 (0.922)
POST-CELEBRITY $> 2h$ (β_3)	-0.177 (0.250)	0.684 (1.262)
Observations	4,748,501	4,748,501
R-squared	0.014	0.045
Time Trend	Linear	Linear
Mean DV	1.08	7.75

The unit of observation is a tweet. The dependent variable is equal to 100 (and zero otherwise) if the tweet contains partisan language (Column 1), or issue polarization (Column 2). The Table shows the OLS estimates of β from Equation (1) for interventions by celebrities (non-political influencers) in the debate. Regressions use all tweets not posted by celebrities; politicians' tweets remain in this sample. Panel A shows the results of β from OLS estimation of Equation (1). Panel B shows the results of β_1 , β_2 , and β_3 from OLS estimation of Equation (2) with $a = b = c = 60$. Standard errors are two-way clustered at the event level and at the level of 30 minutes time intervals since the onset of the debate. Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

To investigate these possibilities, we perform the same analysis for similarly salient, central, public, and, sometimes, divisive, figures, but who are not politicians. We select these influencers through a selection procedure as close as possible to the one used for politicians (see Section 2.5). These “celebrities” intervene in 19% of the debates. They tend to have many more followers than politicians. For example, the 60th member in this list by number of followers (singer songwriter Shawn Mendes) has more than 26 million followers. The analysis on this set of such prominent influencers therefore may provide an upper bound estimate of the effect due to news amplification. Many of the included celebrities, such as Kanye West, Kim Kardashian, or Ellen Degeneres, are controversial and divisive figures, perhaps even more so than some of the politicians included in our list. The estimates should thus also capture effects due to divisiveness.

We estimate Equations (1)–(3) by considering p_e as the timing of the first intervention by a celebrity. Figure A.14 shows the estimates of the event-study specification. The overall net effects in the post-intervention period are undistinguishable from zero, both for partisanship and issue-polarization, as confirmed by the two-way fixed effects results displayed in Table 4. The event-study results show that the levels of partisanship and issue-polarization in the debate remain stable after an intervention by a celebrity. Appendix Table A.15 further shows that, unlike political interventions, interventions by celebrities reduce tribalism overall and in the long run, while they do not consistently shape hate speech. Celebrity interventions also reduce affective polarization overall and in the long run; the medium-term (1–2h) estimate is negative but not statistically significant.

6.2 Political supply of partisan rhetoric polarizes public debates

We now show that politicians polarize the public debate because they supply a rhetoric that casts news events through a partisan lens. In particular, we describe how partisan political interventions have a disproportionately polarizing effect. To show this, we modify Equation (1) to estimate the polarizing effects of interventions with and without partisan or issue-polarizing rhetoric. We estimate:

$$y_{i,e,t} = \beta_S S_{i,e,t}^{(p_e,T)} + \beta_{NS} NS_{i,e,t}^{(p_e,T)} + \eta_e + \theta_t + \zeta t + v_{i,e,t} \quad (5)$$

where $S_{i,e,t}^{(N,M)}$ (respectively $NS_{i,e,t}^{(N,M)}$) takes value 1 between times N and M if the political intervention is of type S (respectively, not of type S), where S measures whether the tweet contains a given rhetoric. We systematically report statistical tests of the difference between coefficients β_S and β_{NS} to capture how the supply of partisanship and issue polarization shapes the public debate. Here, since the language of the tweet may be endogenous, the results are mostly descriptive.

The first panel of Table 5 displays the estimation results (we show in Table A.16 the estimation results of a specification for different post-treatment estimation windows, closer in spirit to Equation 2). Column 1 shows that political interventions that supply partisanship are associated with a 1.46 percentage points increase in the debate’s partisan content after the intervention. Political interventions that do not supply partisan language are also followed by a statistically significant, but smaller, increase of 0.60 percentage points. The difference between the two coefficients (0.86 percentage points) is statistically significant at the 10% level, as shown in the bottom of the panel. Column 2 shows that political interventions that supply partisan language are also associated with a large and statis-

Table 5: Politicians' rhetoric and characteristics and partisanship and issue polarization of the public debate

VARIABLES	(1) Partisanship	(2) Issue polarization	(3) Partisanship	(4) Issue polarization
SUPPLY=0 * POST-Intervention	0.599*** (0.216)	2.012** (0.855)	0.681*** (0.236)	2.644*** (0.905)
SUPPLY=1 * POST-Intervention	1.458*** (0.435)	5.431*** (1.595)	0.792 (0.530)	1.703 (1.655)
Observations	4,747,621	4,747,621	4,747,621	4,747,621
R-squared	0.014	0.045	0.014	0.045
Supply	Partisanship		Issue polarization	
Difference	0.859* (0.499)	3.419* (1.919)	0.112 (0.614)	-0.942 (2.061)
SE				
Type=0 * POST-Intervention	0.592** (0.231)	1.396 (0.855)	0.665** (0.285)	2.444* (1.425)
Type=1 * POST-Intervention	1.093*** (0.253)	5.972*** (1.584)	0.720*** (0.250)	2.431*** (0.906)
Observations	4,747,621	4,747,621	4,747,621	4,747,621
R-squared	0.014	0.045	0.014	0.045
Supply	Republican		Male	
Difference	0.502* (0.291)	4.576*** (1.807)	0.055 (0.377)	-0.014 (1.722)
SE				

The unit of observation is a tweet. The dependent variable is equal to 100 (and zero otherwise) if the tweet contains partisan language (columns 1 and 3), or issue-polarizing language (columns 2 and 4). Panel A shows the results of β_S and β_{NS} from OLS estimation of Equation 5. In columns 1 and 2 of the top panel, S captures whether the political intervention contains partisan language. In columns 3 and 4 of the top panel, S captures whether the political intervention contains issue-polarizing language. In columns 1 and 2 of the bottom panel, S captures whether the political intervention is from a Republican. In columns 3 and 4 of the bottom panel, S captures whether the political intervention is from a male politician. Standard errors are two-way clustered at the event level and at the level of 30 minutes time intervals since the onset of the debate. Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

tically significant 5.43 percentage points increase in the *issue-polarized* content of public debate, compared with a 2.01 percentage point increase for non-partisan interventions, a difference (3.42 percentage points) that is also statistically significant at the 10% level. By contrast, as shown in Columns 3 and 4, the debate is no more partisan or issue-polarized after interventions that supply issue-polarized language than after interventions that do not: the differences between the two coefficients (0.11 and -0.94 percentage points) are small and statistically insignificant. Overall, this evidence suggests that the supply of partisan rhetoric—the unique language of politicians—is key in politicizing online debates through both a partisan and polarized lens.

We investigate the role of politicians' characteristics in the bottom panel of Table 5 (see Table A.17 for the estimation results of a specification with different post-treatment estimation windows). Interventions by Republican politicians are followed by larger increases in both partisan and issue-polarized language in the public debate (columns 1

and 2). The gap is small for partisan phrases (0.50 percentage points, significant at the 10% level) and large for issue-polarized language: the increase is 5.97 percentage points after a Republican intervention, compared with a statistically insignificant 1.40 after a Democratic one (a difference of 4.58 percentage points). Nevertheless, this is not because Republican politicians are more likely than Democrats to be the first to intervene in a debate, nor because they are more likely to use partisan rhetoric compared to Democratic politicians. In fact, Table A.2 shows that Republican politicians are significantly *less* likely to supply partisan rhetoric compared with Democrats. Instead, the difference is explained by differential effects on Democratic- and Republican-leaning partisan phrases, as shown in Appendix Table A.18 and Table A.19. First interventions by a Republican politician lead to an increase in both Democratic- and Republican-leaning partisan phrases, while first interventions by a Democrat only lead to a change in Democratic-leaning partisan phrases. Unlike the results by political affiliation, Columns 3 and 4 in the bottom panel of Table 5 show no statistically significant differences by politicians' gender (the differences are 0.06 and -0.01 percentage points).

6.3 Who is polarized by political interventions?

Finally, we investigate which users are most involved in the online polarization of the debate after a political intervention. First, we study whether political interventions differentially affects users who are already politicized. Second, we study whether political interventions primarily affects users who are already engaged in the debate prior to the intervention (perhaps highly informed users who follow news events more closely) or, instead, *newcomers* to the debate: users who first tweet about the event after the intervention. To do so, we estimate Equation (5) where $S_{i,e,t}^{(N,M)}$ takes value 1 between times N and M if the tweet originates from a "politicized user", which we define as a user who uses political (partisan) identity terms in their profile description.³² To distinguish between users already engaged in the debate and newcomers, we also estimate a version of Equation (5) where $S_{i,e,t}^{(N,M)}$ takes value 1 if the tweet originates from a user who had already tweeted about the debate before the political intervention, and $N S_{i,e,t}^{(N,M)}$ takes value 1 between times N and M if the tweet is from a newcomer to the debate.

Estimation results, displayed in Table A.20, show that political interventions only shape partisanship and issue polarization among newcomers to the debate.³³ While the effects are stronger for already politicized users, particularly for issue-polarization, non-

³²The list of political identity terms consists of terms such as Democrat, Republican, MAGA, socialist. See Appendix B.4.1

³³Returning users and politicized users contribute 16.2% and 11.2% of the tweets in our sample, respectively.

politicized users also adopt a more partisan and issue polarized rhetoric after a political intervention. One possible interpretation of this result would be that political interventions “attract” users to the debate, or that they disproportionately attract more politicized users. Yet, we do not find that political interventions are associated with an increase in the volume of tweets overall (Figure 4) or among politicized or non-politicized users separately (Appendix Figure A.15), nor with an increase in the *share* of tweets by politicized users (Figure A.16).

Therefore, our results are consistent with the idea that the supply of partisan rhetoric by politicians places mass-shooting events under a polarized lens that steers the view and polarizes users who join the debate after the intervention, with a particularly large effect for those who have already formed a political allegiance to one of the parties.

7 Conclusion

One of democracy’s objectives and virtues, as famously discussed by Madison (The Federalist Papers, 10), is to limit unhinged political polarization—the “violence of faction”—by funneling ideological conflicts through institutionalized peaceful debates. Today as in the past, rising political polarization poses a threat to democratic institutions. This paper documents that the polarization of the American political debate is, at least partly, the result of a top-down mechanism, whereby political leaders (whether intentionally or accidentally) poison the public debate. We show this within the context of conversations on social media regarding an important class of salient policy-relevant events: mass shootings. Our results point to a systematic pattern in which politicians, by casting a partisan lens on these events, reshape online discussions towards partisan, issue-polarized, and, eventually, tribal terms.

Our contribution does not necessarily imply that politicians add fuel to the fire of polarization on every topic and for all events. Some topics are likely to become ground of polarized and heated debates without the need for politicians. Others may never become so. Yet, at least for one famously polarized and recurrent topic in the American public debate, politicians play a unique role in politicizing and polarizing the debate along partisan political lines.

Many commentators have speculated that mass political polarization in America is the result of (or even engineered by) choices made by political leaders, and our results are consistent with the view that elite polarization precedes mass polarization (e.g., [Bueno de Mesquita and Dziuda, 2023](#); [Callander and Carbajal, 2022](#); [Canen et al., 2026](#); [Handan-Nader et al., 2026](#); [Phillips et al., 2024](#)). While our data cannot shed light on the extent to which political leaders intentionally engineer political polarization, we provide direct

evidence that their communication efforts cause greater polarization. We speculate that our results are particularly evident in the context of social media precisely because social media facilitate and accelerate top-down mechanisms of polarization, by enabling direct and frequent contact between politicians and citizens. Thus, through this channel, the rise of social media may have naturally coincided with an increase in mass polarization. Normatively, our results also refocus attention from bottom-up causes of polarization (echo-chambers, misinformation, or the flattening of the informational environment) to the role played by elites, and in particular political elites.

References

- Abramowitz, A. I. and S. Webster (2016). The rise of negative partisanship and the nationalization of U.S. elections in the 21st century. *Electoral Studies* 41, 12–22.
- Acemoglu, D., A. Ozdaglar, and A. ParandehGheibi (2010). Spread of (mis)information in social networks. *Games and Economic Behavior* 70(2), 194–227.
- Adena, M., R. Enikolopov, M. Petrova, V. Santarosa, and E. Zhuravskaya (2015). Radio and the rise of the Nazis in prewar Germany. *The Quarterly Journal of Economics* 130(4), 1885–1939.
- Alatas, V., A. G. Chandrasekhar, M. Mobius, B. A. Olken, and C. Paladines (2024). Do celebrity endorsements matter? A Twitter experiment promoting vaccination in Indonesia. *The Economic Journal* 134(659), 913–933.
- Allcott, H., L. Braghieri, S. Eichmeyer, and M. Gentzkow (2020). The welfare effects of social media. *American Economic Review* 110(3), 629–676.
- Amarasinghe, A. and P. A. Raschky (2022). Competing for attention—The effect of talk radio on elections and political polarization in the US. arXiv preprint arXiv:2206.13675.
- Angelucci, C., J. Cagé, and M. Sinkinson (2024). Media competition and news diets. *American Economic Journal: Microeconomics* 16(2), 62–102.
- Angelucci, C. and A. Prat (2024). Is journalistic truth dead? Measuring how informed voters are about political news. *American Economic Review* 114(4), 887–925.
- Anger, I. and C. Kittl (2011). Measuring influence on Twitter. In *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies*, pp. 1–4.
- Ash, E., S. Galletta, M. Pinna, and C. Warshaw (2024). From viewers to voters: Tracing Fox News’ impact on American democracy. *Journal of Public Economics* 240, 105256.
- Ash, E. and S. Hansen (2023). Text algorithms in economics. *Annual Review of Economics* 15(1), 659–688.
- Bakshy, E., J. M. Hofman, W. A. Mason, and D. J. Watts (2011). Everyone’s an influencer: Quantifying influence on Twitter. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, pp. 65–74.
- Bakshy, E., S. Messing, and L. A. Adamic (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science* 348(6239), 1130–1132.
- Barber, M. and N. McCarty (2015). Causes and consequences of polarization. In *Solutions to Political Polarization in America*, pp. 15–58. Cambridge: Cambridge University Press.

- Barilari, F. (2026). Contagious polarization: How divisive shocks reshape legislative behavior across policy domains. Working paper; previously circulated as “Shooting Political Polarization”.
- Beknazar-Yuzbashev, G., R. Jiménez Durán, J. McCrosky, and M. Stalinski (2025). Toxic content and user engagement on social media: Evidence from a field experiment. SSRN working paper.
- Boffa, F., G. Battiston, E. Levi, and S. Stillman (2024). Strategic use of social media by political parties: Evidence from Italy. Working paper, Free University of Bozen/Bolzano.
- Borusyak, K., X. Jaravel, and J. Spiess (2024). Revisiting event-study designs: Robust and efficient estimation. *The Review of Economic Studies* 91(6), 3253–3285.
- Boxell, L., M. Gentzkow, and J. M. Shapiro (2024). Cross-country trends in affective polarization. *The Review of Economics and Statistics* 106(2), 557–565.
- Bueno de Mesquita, E. and W. Dziuda (2023). Partisan traps. NBER Working Paper No. 31827.
- Bursztyn, L., G. Egorov, R. Enikolopov, and M. Petrova (2023). Social media and xenophobia: Evidence from Russia. In F. Campante, R. Durante, and A. Tesei (Eds.), *The Political Economy of Social Media*, pp. 153–161. Paris and London: CEPR Press.
- Bursztyn, L., G. Egorov, and S. Fiorin (2020). From extreme to mainstream: The erosion of social norms. *American Economic Review* 110(11), 3522–3548.
- Bursztyn, L., J. Kolstad, A. Rao, P. Tebaldi, and N. Yuchtman (2025). Polarisation and public policy: Political adverse selection under Obamacare. *The Economic Journal*, ueaf079.
- Cagé, J., N. Hervé, and B. Mazoyer (2022). Social media and newsroom production decisions. CEPR Discussion Paper No. DP17358.
- Callander, S. and J. C. Carbajal (2022). Cause and effect in political polarization: A dynamic analysis. *Journal of Political Economy* 130(4), 825–880.
- Campante, F., R. Durante, and A. Tesei (Eds.) (2023). *The Political Economy of Social Media*. CEPR E-book. CEPR Press.
- Canen, N. J., C. Kendall, and F. Trebbi (2026). Political parties as drivers of U.S. polarization: 1927–2018. *Journal of Political Economy*.
- Couttenier, M., S. Hatte, M. Thoenig, and S. Vlachos (2024). Anti-muslim voting and media coverage of immigrant crimes. *Review of Economics and Statistics* 106(2), 576–585.
- De Chaisemartin, C. and X. d’Haultfoeuille (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review* 110(9), 2964–2996.
- DellaVigna, S., R. Enikolopov, V. Mironova, M. Petrova, and E. Zhuravskaya (2014). Cross-border media and nationalism: Evidence from Serbian radio in Croatia. *American Economic Journal: Applied Economics* 6(3), 103–132.
- DellaVigna, S. and E. Kaplan (2007). The Fox News effect: Media bias and voting. *The Quarterly Journal of Economics* 122(3), 1187–1234.
- Demszky, D., N. Garg, R. Voigt, J. Zou, J. Shapiro, M. Gentzkow, and D. Jurafsky (2019). Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)*, pp. 2970–3005.

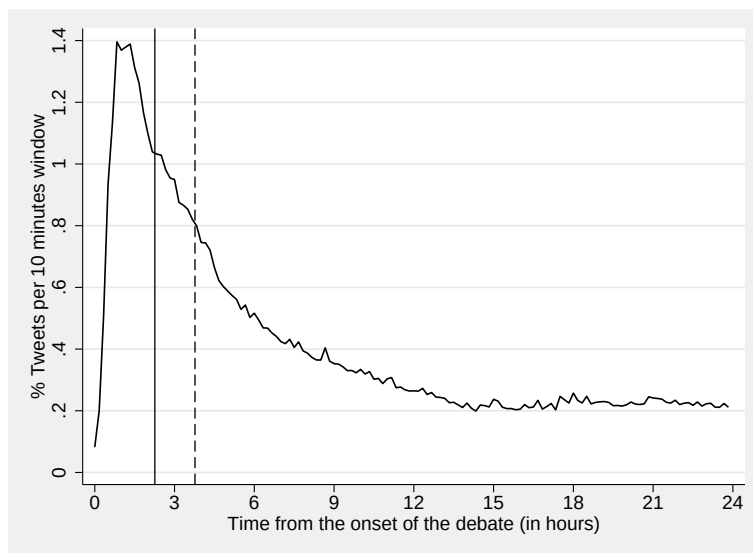
- Djourelouva, M., R. Durante, E. Motte, and E. Patacchini (2024). Experience, narratives, and climate change beliefs. SSRN working paper.
- Ederer, F., P. Goldsmith-Pinkham, and K. Jensen (2025). Anonymity and identity online. Working paper.
- Enikolopov, R., A. Makarin, and M. Petrova (2020). Social media and protest participation: Evidence from Russia. *Econometrica* 88(4), 1479–1514.
- Enke, B. (2020). Moral values and voting. *Journal of Political Economy* 128(10), 3679–3729.
- Enríquez, J. R., H. Larreguy, J. Marshall, and A. Simpser (2024). Mass political information on social media: Facebook ads, electorate saturation, and electoral accountability in Mexico. *Journal of the European Economic Association* 22(4), 1678–1722.
- Fowler, A., S. J. Hill, J. B. Lewis, C. Tausanovitch, L. Vavreck, and C. Warshaw (2023). Moderates. *American Political Science Review* 117(2), 643–660.
- Fujiwara, T., K. Müller, and C. Schwarz (2024). The effect of social media on elections: Evidence from the United States. *Journal of the European Economic Association* 22(3), 1495–1539.
- Funke, M., M. Schularick, and C. Trebesch (2023). Populist leaders and the economy. *American Economic Review* 113(12), 3249–3288.
- Gentzkow, M., J. M. Shapiro, and M. Taddy (2019). Measuring group differences in high-dimensional choices: Method and application to congressional speech. *Econometrica* 87(4), 1307–1340.
- Giavazzi, F., F. Iglhaut, G. Lemoli, and G. Rubera (2024). Terrorist attacks, cultural incidents, and the vote for radical parties: Analyzing text from Twitter. *American Journal of Political Science* 68(3), 1002–1021.
- González-Bailón, S., D. Lazer, P. Barberá, M. Zhang, H. Allcott, T. Brown, A. Crespo-Tenorio, D. Freelon, M. Gentzkow, A. M. Guess, S. Iyengar, Y. M. Kim, N. Malhotra, D. Moehler, B. Nyhan, J. Pan, C. V. Rivera, J. Settle, E. Thorson, R. Tromble, A. Wilkins, M. Wojcieszak, C. K. de Jonge, A. Franco, W. Mason, N. J. Stroud, and J. A. Tucker (2023). Asymmetric ideological segregation in exposure to political news on Facebook. *Science* 381(6656), 392–398.
- Graham, J., J. Haidt, and B. A. Nosek (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology* 96(5), 1029–1046.
- Grosjean, P., F. Masera, and H. Yousaf (2022). Inflammatory political campaigns and racial bias in policing. *The Quarterly Journal of Economics* 138(1), 413–463.
- Guess, A. M., N. Malhotra, J. Pan, P. Barberá, H. Allcott, T. Brown, A. Crespo-Tenorio, D. Dimmery, D. Freelon, M. Gentzkow, S. González-Bailón, E. Kennedy, Y. M. Kim, D. Lazer, D. Moehler, B. Nyhan, C. V. Rivera, J. Settle, D. R. Thomas, E. Thorson, R. Tromble, A. Wilkins, M. Wojcieszak, B. Xiong, C. K. de Jonge, A. Franco, W. Mason, N. J. Stroud, and J. A. Tucker (2023). How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science* 381(6656), 398–404.
- Guess, A. M., B. Nyhan, and J. Reifler (2020). Exposure to untrustworthy websites in the 2016 US election. *Nature Human Behaviour* 4(5), 472–480.
- Guriev, S., E. Henry, T. Marquis, and E. Zhuravskaya (2023). Curtailing false news, amplifying truth. CEPR Discussion Paper No. 18650.

- Guriev, S. and E. Papaioannou (2022). The political economy of populism. *Journal of Economic Literature* 60(3), 753–832.
- Haidt, J. and J. Graham (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research* 20(1), 98–116.
- Halberstam, Y. and B. Knight (2016). Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter. *Journal of Public Economics* 143, 73–88.
- Handan-Nader, C., A. C. W. Myers, and A. B. Hall (2026). Polarization and state legislative elections. *American Journal of Political Science* 70(2), 453–470.
- Hassan, T. A., S. Hollander, A. Kalyani, L. van Lent, M. Schwedeler, and A. Tahoun (2025). Text as data in economic analysis. *Journal of Economic Perspectives* 39(3), 193–220.
- Hatte, S., E. Madinier, and E. Zhuravskaya (2022). Reading Twitter in the newsroom: Web 2.0 and traditional-media reporting of conflicts. CEPR Discussion Paper No. DP16167.
- Henry, E., E. Zhuravskaya, and S. Guriev (2022, August). Checking and sharing alt-facts. *American Economic Journal: Economic Policy* 14(3), 55–86.
- Jackson, M. O., S. Malladi, and D. McAdams (2022). Learning through the grapevine and the impact of the breadth and depth of social networks. *Proceedings of the National Academy of Sciences* 119(34), e2205549119.
- Koleva, S. P., J. Graham, R. Iyer, P. H. Ditto, and J. Haidt (2012). Tracing the threads: How five moral concerns (especially purity) help explain culture war attitudes. *Journal of Research in Personality* 46(2), 184–194.
- Lenz, G. S. and C. Lawson (2011). Looking the part: Television leads less informed citizens to vote based on candidates’ appearance. *American Journal of Political Science* 55(3), 574–589.
- Levy, G. and R. Razin (2019). Echo chambers and their effects on economic and political outcomes. *Annual Review of Economics* 11(1), 303–328.
- Levy, R. (2021, March). Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review* 111(3), 831–870.
- Martin, G. J. and A. Yurukoglu (2017). Bias in cable news: Persuasion and polarization. *American Economic Review* 107(9), 2565–2599.
- Mattozzi, A., S. Nocito, and F. Sobbrío (2023). Fact-checking politicians. SSRN working paper.
- Merod, A. (2016, June 15). How the NRA has responded to mass shootings over the years. NBC News. Accessed September 17, 2025.
- Müller, K. and C. Schwarz (2021). Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association* 19(4), 2131–2167.
- Müller, K. and C. Schwarz (2023). From hashtag to hate crime: Twitter and antiminority sentiment. *American Economic Journal: Applied Economics* 15(3), 270–312.
- Nyhan, B., J. Settle, E. Thorson, M. Wojcieszak, P. Barberá, A. Y. Chen, H. Allcott, T. Brown, A. Crespo-Tenorio, D. Dimmery, D. Freelon, M. Gentzkow, S. González-Bailón, A. M. Guess, E. Kennedy, Y. M. Kim, D. Lazer, N. Malhotra, D. Moehler, J. Pan, D. R. Thomas, R. Tromble, C. V. Rivera, A. Wilkins, B. Xiong, C. K. de Jonge, A. Franco, W. Mason, N. J. Stroud, and J. A. Tucker (2023). Like-minded sources on Facebook are prevalent but not polarizing. *Nature* 620(7972), 137–144.

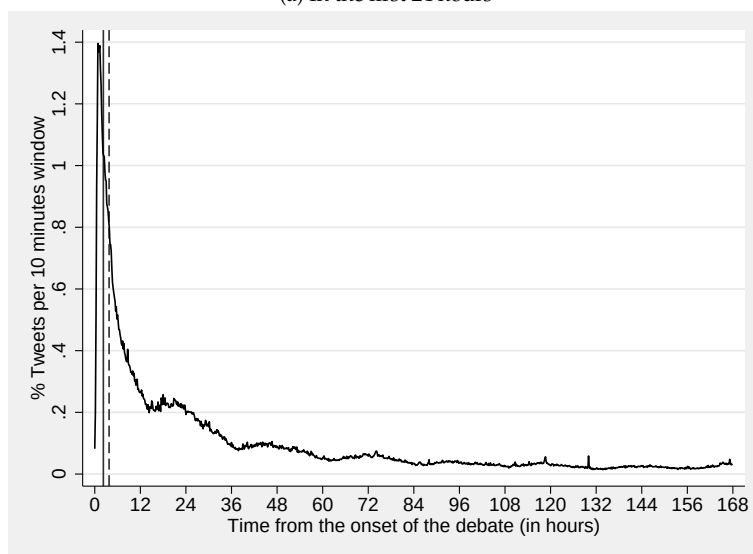
- Phillips, C. H., J. M. Snyder, Jr., and A. B. Hall (2024). Who runs for Congress? A study of state legislators and congressional polarization. *Quarterly Journal of Political Science* 19(1), 1–25.
- Qin, B., D. Strömberg, and Y. Wu (2024). Social media and collective action in China. *Econometrica* 92(6), 1993–2026.
- Rambachan, A. and J. Roth (2023). A more credible approach to parallel trends. *Review of Economic Studies* 90(5), 2555–2591.
- Roth, J. (2022). Pretest with caution: Event-study estimates after testing for parallel trends. *American Economic Review: Insights* 4(3), 305–322.
- Siegel, A. A., E. Nikitin, P. Barberá, J. Sterling, B. Pullen, R. Bonneau, J. Nagler, and J. A. Tucker (2021). Trumping hate on Twitter? Online hate speech in the 2016 US election campaign and its aftermath. *Quarterly Journal of Political Science* 16(1), 71–104.
- Simchon, A., W. J. Brady, and J. J. Van Bavel (2022, 03). Troll and divide: The language of online polarization. *PNAS Nexus* 1(1), pgac019.
- Spenkuch, J. L., E. Teso, and G. Xu (2023). Ideology and performance in public organizations. *Econometrica* 91(4), 1171–1203.
- Sun, L. and S. Abraham (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics* 225(2), 175–199.
- Thorp, H. H. and V. Vinson (2024). Context matters in social media. *Science* 385(6716), 1393–1393.
- Wang, T. (2021a). Media, pulpit, and populist persuasion: Evidence from Father Coughlin. *American Economic Review* 111(9), 3064–3092.
- Wang, T. (2021b). Waves of empowerment: Black radio and the Civil Rights Movement. Working paper.
- Wang, T. (2025). The electric telegraph, news coverage, and political participation. *The Journal of Economic History* 85(1), 1–32.
- Yanagizawa-Drott, D. (2014). Propaganda and conflict: Evidence from the Rwandan genocide. *The Quarterly Journal of Economics* 129(4), 1947–1994.
- Yousaf, H. (2021). Sticking to one’s guns: Mass shootings and the political economy of gun control in the United States. *Journal of the European Economic Association* 19(5), 2765–2802.
- Zhuravskaya, E., M. Petrova, and R. Enikolopov (2020). Political effects of the internet and social media. *Annual Review of Economics* 12, 415–438.

A Appendix: Additional Results and Robustness

A.1 Characteristics of political interventions



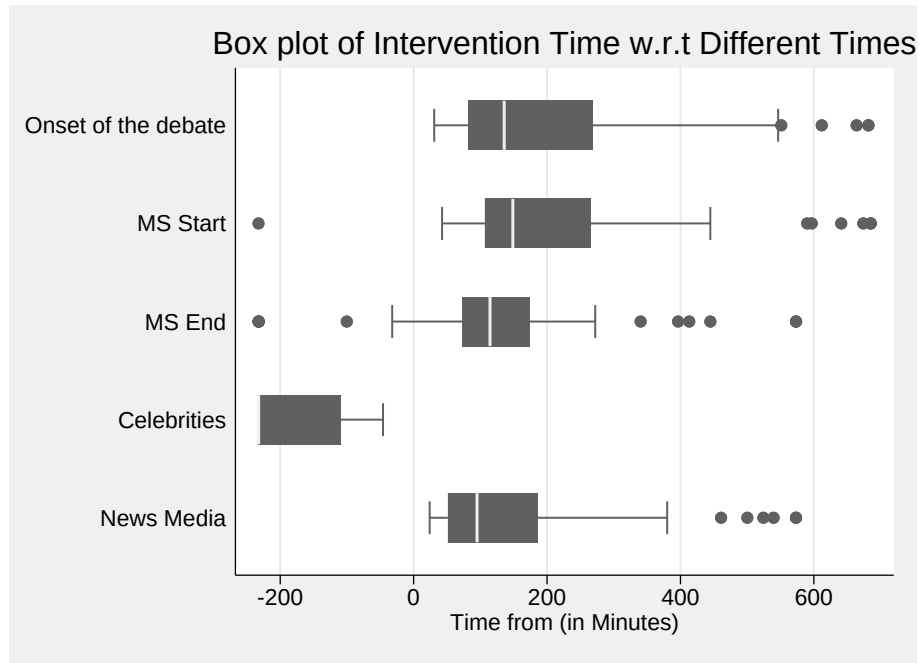
(a) In the first 24 hours



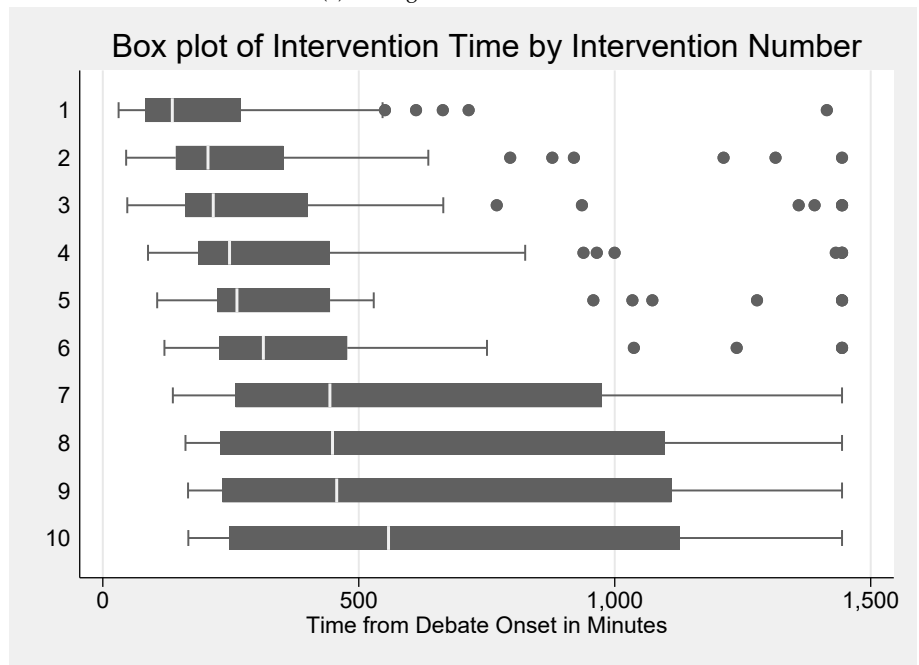
(b) In the entire sample

Figure A.1: Distribution of tweets and the timing of political interventions

These figures plot the distribution of tweets in each 10 minute bin from the onset of the debate. The y-axis plots the percentage of tweets that occur in each 10 minute bin. The x-axis plots the time from the onset of the debate (in hours). In panel (a) the figure zooms in the first 24 hours from the onset of the debate. In panel (b) the figure shows the distribution of tweets for the entire debate, the first 7 days after the onset of the debate. In each panel, the solid vertical line indicates the median intervention time by a politician, while the dashed line indicates the average intervention time by a politician.



(a) Timing of first intervention



(b) Timing of subsequent interventions

Figure A.2: Timing of political interventions

These figures present box plots of the timing of political interventions. Panel (a) plots the 25th percentile, median, 75th percentile, and 1.5 times the inter-quartile range (boxplot) of the timing of the first intervention w.r.t. five key temporal reference points of an event: the onset of the public debate, the start of the mass shooting event, the end of the mass shooting event (arrest, incapacitation, or death of the shooter), and the first intervention in the public debate by a celebrity. Panel (b) displays a boxplot showing the timing of the first ten political interventions relative to the onset of the public debate. All timing variables are winsorized at the 95th percentile.

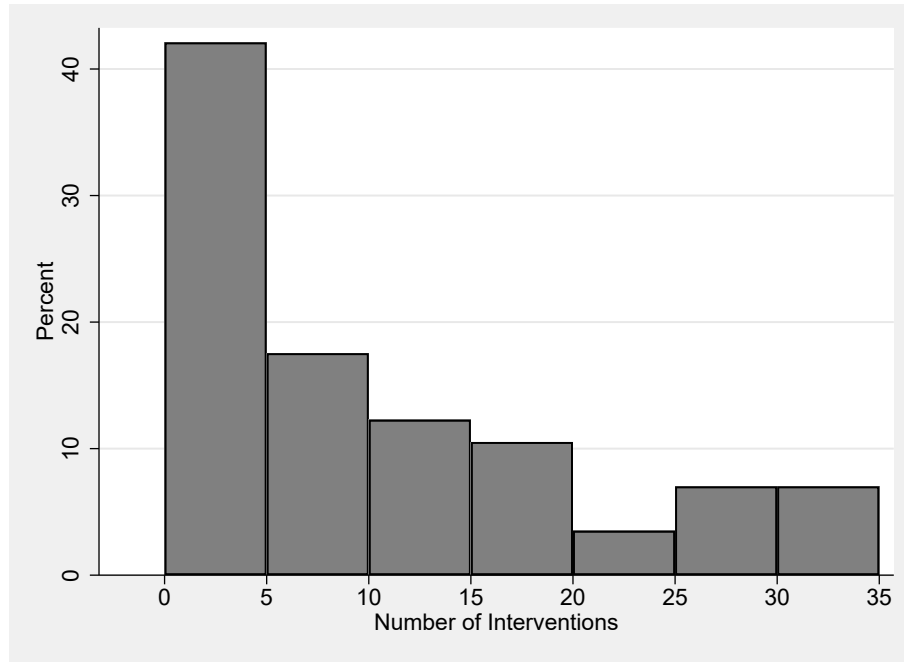


Figure A.3: Number of political interventions

The figure shows the distribution (histogram) of the number of political interventions in a public debate.

Table A.1: Number of political interventions

Dependent Variable (DV): Number of Interventions			
Independent Variable (X)	Mean of DV when:		Difference
	X=0	X=1	
Panel A: Politician characteristics			
Male	6.88 (5.00)	13.06 (11.21)	6.17*** (2.25)
Republican	8.59 (8.48)	18.38 (10.97)	9.79*** (3.28)
SUPPLY: Partisanship	11.74 (10.53)	8.10 (7.13)	-3.64 (2.73)
SUPPLY: Issue polarization	9.97 (8.69)	14.23 (13.12)	4.26 (3.83)
SUPPLY: Tribalism	10.83 (9.92)	11.30 (10.35)	0.48 (2.83)
SUPPLY: Affective polarization	9.96 (8.55)	12.04 (11.27)	2.08 (2.76)
Followers Count: High	10.53 (10.7)	12.19 (8.46)	1.66 (2.75)
Panel B: Event characteristics			
Shooter race: White	9.18 (8.33)	13.04 (11.14)	3.85 (2.75)
Shooter race: Black	12.64 (10.83)	6.73 (4.82)	-5.91** (2.25)
Shooter race: Other	11.26 (10.12)	11.64 (10.45)	0.38 (3.47)
Shooting location: School	11.54 (9.93)	10.17 (10.73)	-1.37 (3.42)
Shooting location: Business	10.21 (9.65)	15.89 (11.01)	5.67 (3.83)
Shooting location: Community	11.02 (9.83)	12.25 (11.72)	1.23 (4.23)
Shooting deaths: High	6.54 (6.18)	20.29 (10.16)	13.75*** (2.65)
Shooting length: Long	11.24 (10.42)	12.88 (8.9)	1.63 (3.43)
Panel C: Twitter characteristics before intervention			
Tweets Volume: High	11.05 (9.99)	11 (10.41)	-0.05 (3.12)
Posters' Followers: High	10.66 (10.42)	11.65 (9.55)	0.99 (2.82)
Posters' Followings: High	9.79 (9.42)	13.21 (10.89)	3.42 (2.98)
Event's Like Count: High	9.94 (9.4)	13.5 (11.18)	3.56 (3.18)
Event's Retweet Count: High	10.42 (9.33)	11.95 (11.12)	1.53 (2.94)

The unit of observation is an event. Column 1 (respectively, 2) shows the mean and standard deviation of the number of political interventions when the variable in the corresponding row is equal to 0 (respectively, 1). Column 3 displays the coefficients estimated from separate OLS regressions (with robust standard errors) of the number of political interventions (dependent variable) on each row variable (independent variable). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.2: Partisan and issue-polarized language in interventions, by account type

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Partianship	Issue polarization	Partianship	Issue polarization	Partianship	Issue polarization	Partianship	Issue polarization
Politicians	17.686*** (5.374)	16.138*** (5.910)			18.840*** (1.327)	24.393*** (1.558)		
News Media	-0.882*** (0.102)	-4.060*** (0.375)			1.384*** (0.089)	5.196*** (0.194)		
Celebrities	-1.182*** (0.019)	9.791 (11.629)			2.267 (3.388)	5.403 (6.404)		
Republican			-19.535** (7.634)	1.257 (15.419)			-5.153* (2.792)	-4.884 (3.295)
Male			-18.317 (13.656)	19.307 (12.011)			-8.673** (4.353)	-3.749 (4.696)
Observations	324,967	324,967	52	52	355,968	355,968	909	909
R-squared	0.001	0.000	0.122	0.046	0.008	0.005	0.013	0.004
Mean Supply	1.18	8.36	19.23	25.00	1.36	8.94	20.02	32.78
Intervention	First	First	First	First	All	All	All	All
Politicians - News Media	18.568*** (5.375)	20.198 (13.045)			17.456*** (3.639)	19.197*** (1.568)		
Politicians - Celebrities	18.868*** (5.374)	6.346 (5.922)			16.574*** (1.330)	18.990*** (6.590)		

The dependent variable is equal to 100 (and zero otherwise) if the tweet contains partisan language (odd columns), or issue-polarizing language (even columns). The unit of observation is a tweet in Columns 1, 2, 5, 6 or a tweet by a politician in Columns 3, 4, 7, 8. Columns 1 and 2 (and, respectively, columns 5 and 6) use a sample comprising all tweets posted before the first political intervention, together with the first (respectively, all) interventions by politicians, celebrities, or news media. Columns 3 and 4 (and, respectively, columns 7 and 8) restrict the sample to the first (respectively, all) interventions by politicians. *** p<0.01, ** p<0.05, * p<0.1.

Table A.3: Tribalism, affective polarization, and hate speech in interventions, by account type

VARIABLES	(1) Tribalism	(2) Affective polarization	(3) Hate Speech	(4) Tribalism	(5) Affective polarization	(6) Hate Speech	(7) Tribalism	(8) Affective polarization	(9) Hate Speech	(10) Tribalism	(11) Affective polarization	(12) Hate Speech
Political	35.843*** (6.838)	13.227* (6.858)	-0.592 (4.353)				43.915*** (1.655)	20.133*** (1.629)	13.500*** (1.445)			
Celebrities	54.197*** (14.504)	5.852 (15.013)	6.269 (11.629)				49.181*** (9.146)	22.466** (9.011)	26.019*** (9.010)			
News	-4.010*** (0.417)	4.967*** (0.911)	-1.219** (0.567)				5.469*** (0.202)	3.190*** (0.285)	12.486*** (0.243)			
Republican				7.597 (17.534)	-16.051 (17.010)	13.899 (12.906)				-2.413 (3.478)	-7.858** (3.401)	22.535*** (2.991)
Male				-15.099 (15.985)	-5.693 (15.585)	4.455 (8.191)				3.670 (4.792)	8.347* (4.748)	7.702** (3.223)
Observations	324,967	324,967	324,967	52	52	52	355,968	355,968	355,968	909	909	909
R-squared	0.001	0.000	0.000	0.019	0.027	0.047	0.008	0.001	0.012	0.001	0.007	0.080
Mean Supply	9.41	39.65	11.90	44.23	51.92	11.54	10.06	39.95	13.11	53.36	59.74	25.41
Intervention	First	First	First	First	First	First	All	All	All	All	All	All
Politicians - Celebrities	-18.353	7.376	-6.861				-5.265	-2.333	-12.518			
SE	16.035	6.917	4.389				1.666	1.649	9.125			
Politicians - News Media	39.853	8.260	0.628				38.447	16.943	1.014			
SE	6.850	16.505	12.417				9.294	9.156	1.463			
Mean Male				0.67	0.67	0.67				0.84	0.84	0.84
Mean Republican				0.25	0.25	0.25				0.45	0.45	0.45

The unit of observation is a tweet (Columns 1–3 and 7–9) or a tweet by a politician (Columns 4–6 and 10–12). The dependent variable is equal to 100 (and zero otherwise) if the tweet contains tribal language (Columns 1, 4, 7, and 10), affective-polarizing language (Columns 2, 5, 8, and 11), or hate speech (Columns 3, 6, 9, and 12). Columns 1 to 3 (and, respectively, columns 7 to 9) use a sample comprising all tweets posted before the first political intervention, together with the first (respectively, all) interventions by politicians, celebrities, or news media. Columns 4 to 6 (and, respectively, columns 10 to 12) restrict the sample to the first (respectively, all) interventions by politicians. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.4: Most frequent dictionary keywords, by account type

Keyword	Dictionary	Tweets containing keyword (%)			
		Regular users	Politicians	Celebrities	News media
law	Issue polarization	3.444	10.121 ⁺	3.448	3.824 ⁺
dem	Issue polarization	3.341	10.341 ⁺	3.448	2.557 ⁻
nra	Issue polarization	2.360	2.750	0.000	0.830 ⁻
polit	Issue polarization	2.298	0.770 ⁻	0.000	1.053 ⁻
gun control	Issue polarization	1.913	0.880 ⁻	0.000	1.675 ⁻
guncontrol	Issue polarization	1.625	0.220 ⁻	3.448	0.024 ⁻
gop	Issue polarization	1.607	1.430	0.000	0.779 ⁻
repub	Issue polarization	1.299	2.090	0.000	0.718 ⁻
republican	Issue polarization	1.233	2.090 ⁺	0.000	0.712 ⁻
democrat	Issue polarization	1.165	0.880	0.000	0.872 ⁻
gun violenc	Partisanship	1.048	18.592 ⁺	3.448	0.969
senat	Issue polarization	0.934	5.831 ⁺	0.000	0.809 ⁻
lib	Issue polarization	0.874	0.880	0.000	0.248 ⁻
social	Issue polarization	0.797	0.220	3.448	0.857
hate crime	Partisanship	0.783	1.210	0.000	0.999 ⁺

Notes: Entries are percentages of tweets in each group containing the keyword; a tweet contributes at most once to a keyword's frequency. The table reports the 15 keywords with the highest frequency among regular-user tweets. Regular users exclude politicians, celebrities, and news media. Group sizes are 4,714,463 regular-user tweets, 909 politician tweets, 29 celebrity tweets, and 33,129 news-media tweets. Relative to regular users, ⁺ (⁻) denotes a significantly higher (lower) frequency at the 5% level, based on a two-sided Fisher exact test. Tests are unadjusted for multiple comparisons.

A.2 Alternative measurement and content of the public debate

Table A.5: First political intervention and model-based issue polarization

VARIABLES	(1) Issue polarization (model-based)	(2) Left issue polarized (model-based)	(3) Right issue polarized (model-based)
Equation 1			
POST-Intervention (β)	2.139** (1.045)	1.139* (0.650)	0.863* (0.515)
Observations	4,747,621	4,747,621	4,747,621
R-squared	0.094	0.079	0.027
Time Trend	Linear	Linear	Linear
Equation 2			
POST-Intervention $< 1h$ (β_1)	0.981 (0.644)	0.331 (0.461)	0.584* (0.292)
POST-Intervention $1 - 2h$ (β_2)	3.021** (1.175)	1.694** (0.773)	1.142** (0.526)
POST-Intervention $> 2h$ (β_3)	3.818* (1.981)	2.372* (1.239)	1.199 (1.093)
Observations	4,747,621	4,747,621	4,747,621
R-squared	0.094	0.079	0.027
Time Trend	Linear	Linear	Linear
Mean DV	8.97	5.92	2.34

The unit of observation is a tweet. The dependent variable is equal to 100 (and zero otherwise) if the fine-tuned DeBERTa classifier codes the tweet as issue-polarized (Column 1), US left/liberal-aligned issue-polarized (Column 2), or US right/conservative-aligned issue-polarized (Column 3); see Appendix B.3 for the construction of the measure. Panel A ('Equation 1') shows the OLS estimate of β from Equation (1). Panel B ('Equation 2') shows the OLS estimates of β_1 , β_2 , and β_3 from Equation (2) with $a = b = c = 60$. Standard errors are two-way clustered at the event level and at the level of 30-minute time intervals since the onset of the debate. Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.6: First political intervention and the prevalence of individual dictionary terms

Panel A: Issue polarization dictionary										
VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	law	nra	guncontrol	dem	gop	gun control	polit	repub	republican	lib
POST-Intervention (β)	1.269*** (0.422)	0.067 (0.318)	0.668* (0.364)	-0.087 (0.192)	0.116** (0.044)	0.723*** (0.197)	-0.007 (0.176)	0.062 (0.053)	0.052 (0.056)	0.123* (0.073)
Observations	4,747,621	4,747,621	4,747,621	4,747,621	4,747,621	4,747,621	4,747,621	4,747,621	4,747,621	4,747,621
R-squared	0.009	0.022	0.019	0.013	0.013	0.007	0.008	0.012	0.012	0.003
Mean DV	1.942	1.124	0.903	0.849	0.820	0.669	0.575	0.476	0.440	0.306

Panel B: Partisanship dictionary										
VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	gun violenc	hate crime	mental health	radic islam	american peopl	public health	middl class	god pleas	illeg immigr	busi owner
POST-Intervention (β)	0.387*** (0.096)	0.206 (0.162)	0.065 (0.058)	-0.012 (0.015)	-0.001 (0.021)	-0.005 (0.009)	-0.009* (0.005)	0.008 (0.005)	0.011 (0.007)	0.013* (0.007)
Observations	4,747,621	4,747,621	4,747,621	4,747,621	4,747,621	4,747,621	4,747,621	4,747,621	4,747,621	4,747,621
R-squared	0.006	0.034	0.005	0.004	0.001	0.000	0.000	0.000	0.008	0.001
Mean DV	0.700	0.123	0.107	0.023	0.022	0.022	0.019	0.015	0.011	0.003

The unit of observation is a tweet. The dependent variable is equal to 100 (and zero otherwise) if the tweet contains the dictionary word indicated in the column heading. Panel A reports results for the ten most frequent words in the issue-polarization dictionary; Panel B reports results for the ten most frequent words in the partisanship dictionary. Each column shows the OLS estimate of β from Equation (1). Standard errors are two-way clustered at the event level and at the level of 30 minutes time intervals since the onset of the debate. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table A.7: First political intervention and the issue domain of the debate

VARIABLES	(1) None	(2) Gun law	(3) Mental health	(4) Blame media	(5) Crime	(6) Race	(7) Gender	(8) School policy	(9) Terrorism	(10) Immigration
POST-Intervention (β)	-3.714** (1.521)	3.853*** (0.842)	0.228 (0.200)	-0.778 (0.677)	0.467 (0.908)	0.247 (0.515)	0.558 (0.481)	-0.043 (0.151)	-0.590 (0.651)	-0.147 (0.150)
Observations	4,747,621	4,747,621	4,747,621	4,747,621	4,747,621	4,747,621	4,747,621	4,747,621	4,747,621	4,747,621
R-squared	0.116	0.064	0.007	0.043	0.104	0.175	0.048	0.045	0.069	0.020
Time Trend	Linear	Linear	Linear	Linear	Linear	Linear	Linear	Linear	Linear	Linear
Equation	1	1	1	1	1	1	1	1	1	1

The unit of observation is a tweet. In each column the dependent variable is equal to 100, and zero otherwise, if the DeBERTa classifier assigns the tweet to the corresponding issue domain, or to no issue domain, column 1, "None". Each column is a separate OLS regression reporting the estimate of β from Equation (1). Standard errors are two-way clustered at the event level and at the level of 30-minute time intervals since the onset of the debate. Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

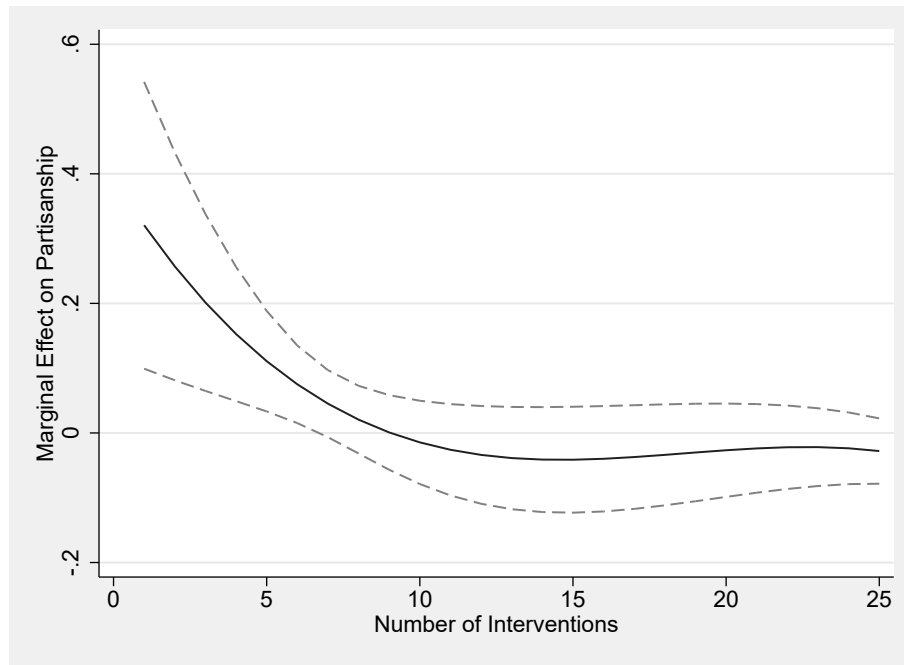
X

Table A.8: First political intervention and issue polarization, by issue domain

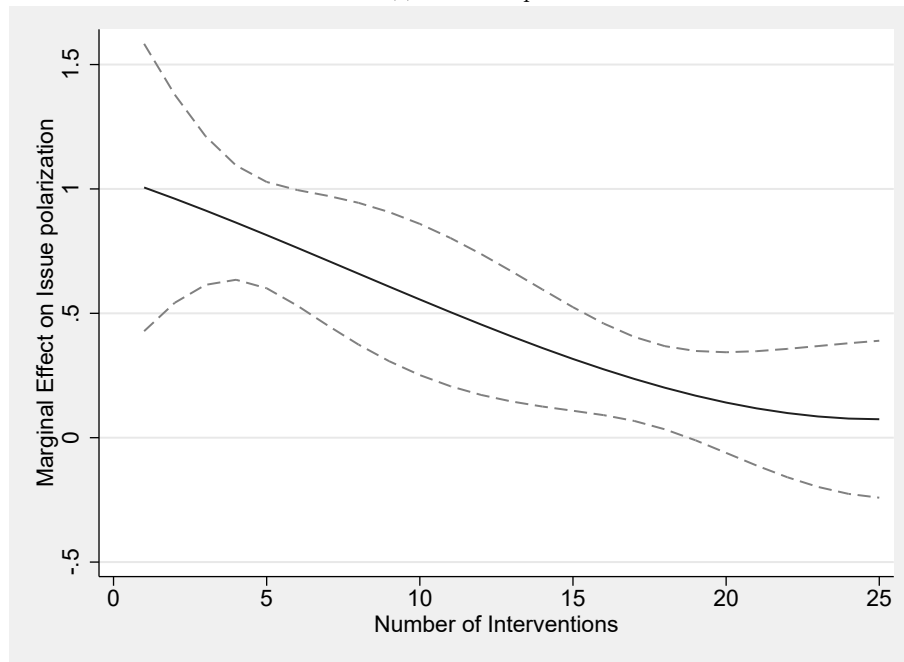
VARIABLES	(1) Issue Polarization	(2) Issue Polarization	(3) Issue Polarization	(4) Issue Polarization	(5) Issue Polarization	(6) Issue Polarization	(7) Issue Polarization	(8) Issue Polarization	(9) Issue Polarization
POST-Intervention (β)	-2.755** (1.147)	-2.838 (4.885)	6.868** (3.187)	-0.567 (2.385)	6.780*** (2.404)	-5.997 (5.164)	1.041 (2.024)	12.082*** (2.650)	-5.325 (3.944)
Observations	735,127	48,075	307,384	195,830	288,123	80,372	75,150	233,351	45,781
R-squared	0.042	0.131	0.065	0.109	0.097	0.071	0.083	0.050	0.067
Domain	Gun law	Mental health	Elite media blame	Crime policing	Race hate crime	Gender LGBT	School policy	Terror national security	Immigration
Time Trend	Linear	Linear	Linear	Linear	Linear	Linear	Linear	Linear	Linear
Equation	1	1	1	1	1	1	1	1	1

The unit of observation is a tweet. Each column reports a separate OLS regression of Equation (1), estimated on the subsample of tweets that the DeBERTa classifier assigns to the issue domain named in the "Domain" row. The dependent variable is equal to 100, and zero otherwise, if the tweet is issue-polarized (model-based). The "none" domain is omitted because such tweets are unpolarized by construction. Standard errors are two-way clustered at the event level and at the level of 30-minute time intervals since the onset of the debate. Robust standard errors are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

A.3 Intensive margin results



(a) Partisanship



(b) Issue polarization

Figure A.4: Number of political interventions: intensive margin results

These figures plot the marginal effect with 95% confidence intervals from OLS estimation of the dependent variable as a quartic function of the number of political interventions. In panel (a), the dependent variable is Partisanship; in panel (b), the outcome variable is Issue Polarization. Standard errors are two-way clustered at the event level and at the level of 30 minutes time intervals since the onset of the debate.

A.4 Identification and robustness

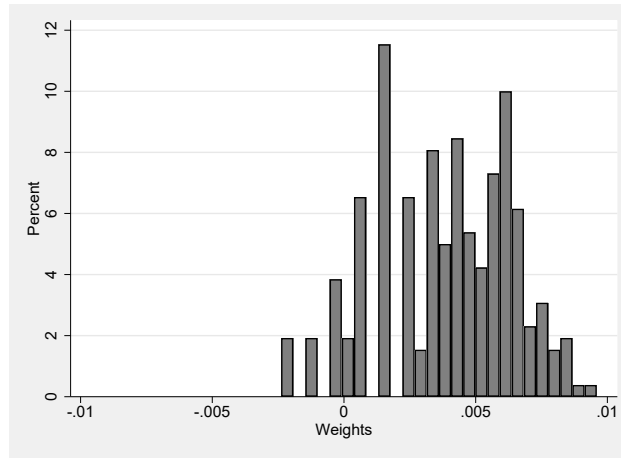
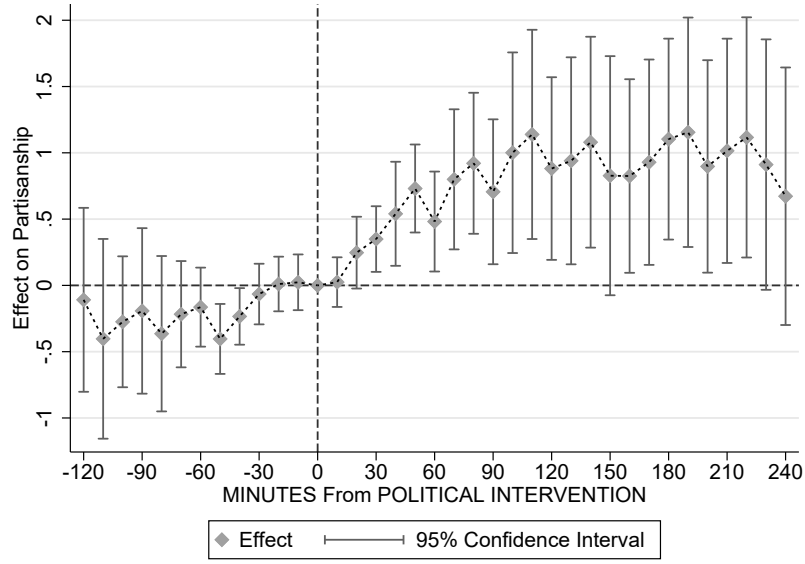
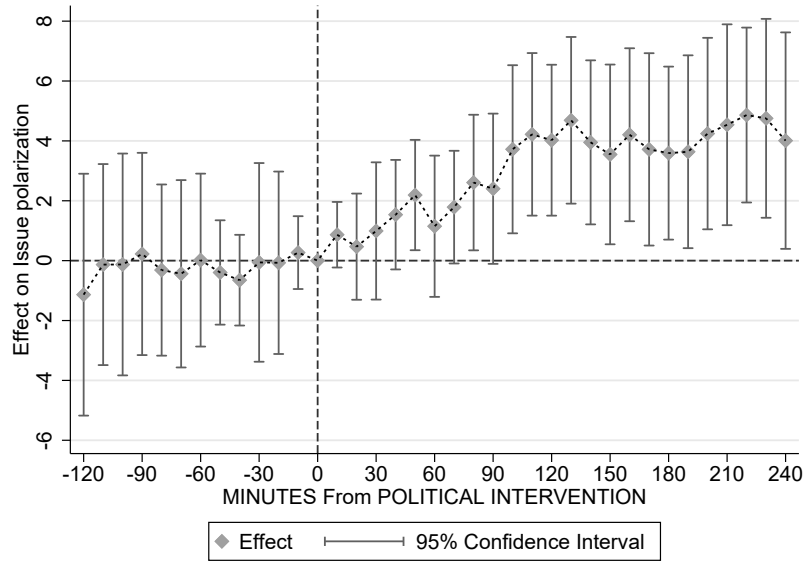


Figure A.5: Difference-in-differences weights following de Chaisemartin and D’Haultfoeuille (2020)

The figure shows the differences-in-differences weight computed using [De Chaisemartin and d’Haultfoeuille \(2020\)](#). The figure focuses on the short and medium-term impact of treatment. For calculating the weights, the data is converted into 30-minute intervals and restricted to 120 minutes from the onset of the treatment.



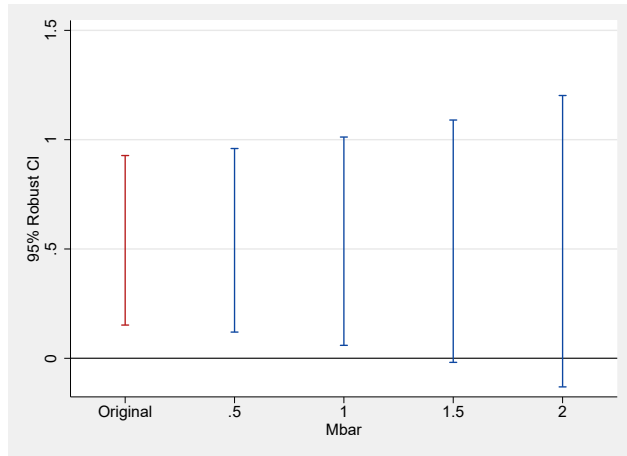
(a) Partisanship



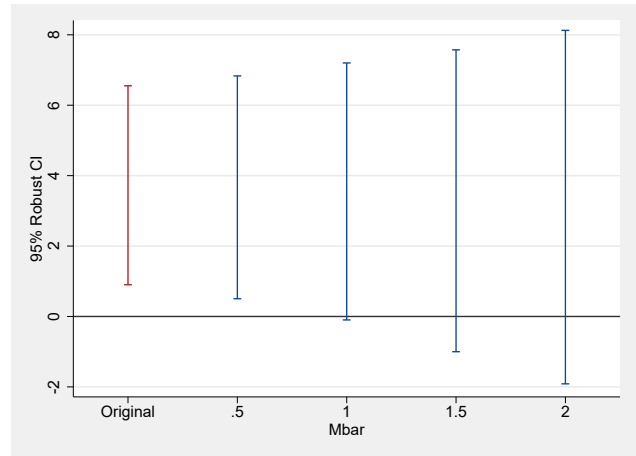
(b) Issue polarization

Figure A.6: Event-study results with no time trend

These figures plot OLS coefficients with 95% confidence intervals. The plotted coefficients are the β_τ coefficients associated with each 10 minute window from Equation (3), when we drop the time trend from Equation (3). The outcome variable is Partisanship in panel (a) and Issue Polarization in panel (b). Standard errors are two-way clustered at the event level and at the level of 30 minutes time intervals since the onset of the debate.



(a) Partisanship



(b) Issue polarization

Figure A.7: First political intervention, partisanship and issue polarization: Robustness to parallel trends violations

The Figure displays robust confidence sets for β from Equation (1) for different values for the parameter $\bar{M}$ described in [Rambachan and Roth \(2023\)](#), which bound the maximal post-treatment violation of parallel trends by $\bar{M}$ times the maximal pre-treatment violation of parallel trends.

Table A.9: First political intervention, partisanship and issue polarization: Robustness to controlling separately for tweets within 1- or 2-min post-intervention buffers

VARIABLES	(1) Partisanship	(2) Issue polarization	(3) Partisanship	(4) Issue polarization
Equation 1				
POST-Intervention (β)	0.701*** (0.198)	2.430*** (0.752)	0.698*** (0.197)	2.426*** (0.750)
Observations	4,747,621	4,747,621	4,747,621	4,747,621
R-squared	0.014	0.045	0.014	0.045
Time Trend	Linear	Linear	Linear	Linear
Equation	1	1	1	1
Buffer in minutes	1	1	2	2
Equation 2				
POST-Intervention<1h (β_1)	0.464*** (0.120)	1.313*** (0.383)	0.463*** (0.120)	1.314*** (0.382)
POST-Intervention 1-2h (β_2)	0.952*** (0.302)	3.122*** (0.808)	0.948*** (0.301)	3.126*** (0.807)
POST-Intervention>2h (β_3)	0.977*** (0.360)	4.229*** (1.369)	0.973*** (0.359)	4.233*** (1.368)
Observations	4,747,621	4,747,621	4,747,621	4,747,621
R-squared	0.014	0.045	0.014	0.045
Time Trend	Linear	Linear	Linear	Linear
Equation	2	2	2	2
Mean DV	1.08	7.75	1.08	7.75
Buffer in minutes	1	1	2	2

The unit of observation is a tweet. The dependent variable is equal to 100 (and zero otherwise) if the tweet contains partisan language (Columns 1 and 3), or issue-polarizing language (Columns 2 and 4). Panel A shows the OLS estimates of β from Equation (1), controlling separately for tweets posted less than one or two minutes after the political intervention (as indicated). Panel B shows the OLS estimates of β_1 , β_2 , and β_3 from Equation (2) with $a = b = c = 60$, controlling separately for tweets posted less than one (Columns 1 and 2) or two (Columns 3 and 4) minutes after the political intervention (as indicated). Standard errors are two-way clustered at the event level and at the level of 30 minutes time intervals since the onset of the debate. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.10: First political intervention, partisanship and issue polarization: Robustness to event-specific time trends

VARIABLES	(1) Partisanship	(2) Issue polarization
Equation 1		
POST-Intervention (β)	0.543*** (0.192)	2.066*** (0.739)
Observations	4,747,621	4,747,621
R-squared	0.015	0.047
Time Trend	Event $\times$ Linear	Event $\times$ Linear
Equation 2		
POST-Intervention<1h (β_1)	0.372*** (0.118)	1.169*** (0.422)
POST-Intervention 1-2h (β_2)	0.786** (0.300)	2.723*** (0.791)
POST-Intervention>2h (β_3)	0.718* (0.380)	3.809** (1.458)
Observations	4,747,621	4,747,621
R-squared	0.015	0.047
Time Trend	Event $\times$ Linear	Event $\times$ Linear
Mean DV	1.08	7.75

The unit of observation is a tweet. The dependent variable is equal to 100 (and zero otherwise) if the tweet contains partisan language (column 1), or issue-polarizing language (column 2). Panel A shows the OLS estimates of β from Equation (1), with the addition of event-specific time trends. Panel B shows the OLS estimates of β_1 , β_2 , and β_3 from Equation (2) with $a = b = c = 60$ and with the addition of event-specific time trends. Standard errors are two-way clustered at the event level and at the level of 30 minutes time intervals since the onset of the debate. Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.11: First political intervention, partisanship and issue polarization: Robustness to alternative definitions of time fixed effects

VARIABLES	(1) Partisanship	(2) Issue polarization	(3) Partisanship	(4) Issue polarization
Equation 1				
POST-Intervention (β)	0.703*** (0.201)	2.387*** (0.762)	0.719*** (0.197)	2.526*** (0.716)
Observations	4,747,621	4,747,621	4,747,621	4,747,621
R-squared	0.014	0.045	0.014	0.045
Time Trend	Linear	Linear	Linear	Linear
Equation	1	1	1	1
FE Window in minutes	15	15	60	60
Equation 2				
POST-Intervention<1h (β_1)	0.462*** (0.121)	1.243*** (0.371)	0.484*** (0.114)	1.430*** (0.519)
POST-Intervention 1-2h (β_2)	0.960*** (0.304)	3.098*** (0.815)	0.953*** (0.279)	3.209*** (0.780)
POST-Intervention>2h (β_3)	0.972** (0.359)	4.175*** (1.377)	1.000*** (0.356)	4.262*** (1.400)
Observations	4,747,621	4,747,621	4,747,621	4,747,621
R-squared	0.014	0.045	0.014	0.045
Time Trend	Linear	Linear	Linear	Linear
Equation	2	2	2	2
Mean DV	1.08	7.75	1.08	7.75
FE Window in minutes	15	15	60	60

The unit of observation is a tweet. The dependent variable is equal to 100 (and zero otherwise) if the tweet contains partisan language (columns 1 and 3), or issue-polarizing language (columns 2 and 4). Panel A shows the OLS estimates of β from Equation (1) with θ_t defined as either 15- (columns 1 and 2) or 60-minutes (columns 3 and 4) time intervals since the onset of the debate (instead of 30 minutes as in our baseline specification). Panel B shows the OLS estimates of β_1 , β_2 , and β_3 from Equation (2) with $a = b = c = 60$ and with θ_t defined as either 15- or 60-minutes time intervals since the onset of the debate (instead of 30 minutes as in our baseline specification). Standard errors are two-way clustered at the event level and at the level of 30 minutes time intervals since the onset of the debate. Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.12: First political intervention, partisanship and issue polarization: Robustness to changing the set of politicians

VARIABLES	(1) Partisanship	(2) Issue polarization	(3) Partisanship	(4) Issue polarization
POST-Intervention (β)	0.831*** (0.218)	2.577*** (0.835)	0.756*** (0.188)	2.219** (0.891)
Observations	4,747,157	4,747,157	4,746,244	4,746,244
R-squared	0.014	0.045	0.014	0.045
Time Trend	Linear	Linear	Linear	Linear
Equation	1	1	1	1
Politicians with followers	>500k	>500k	>200k	>200k
Equation 2				
POST-Intervention<1h (β_1)	0.552*** (0.159)	1.322*** (0.450)	0.456*** (0.107)	0.666 (0.472)
POST-Intervention 1-2h (β_2)	1.123*** (0.321)	3.507*** (0.740)	0.866*** (0.231)	3.261*** (0.796)
POST-Intervention>2h (β_3)	1.229*** (0.364)	4.805*** (1.288)	1.300*** (0.321)	4.692*** (1.234)
Observations	4,747,157	4,747,157	4,746,244	4,746,244
R-squared	0.014	0.045	0.014	0.045
Time Trend	Linear	Linear	Linear	Linear
Equation	2	2	2	2
Mean DV	0.99	7.22	0.88	6.40
Politicians with followers	>500k	>500k	>200k	>200k

The unit of observation is a tweet. We now consider the first political intervention by any politician with at least 500,000 (Columns 1 and 2) or 200,000 followers (Columns 3 and 4). The dependent variable is equal to 100 (and zero otherwise) if the tweet contains partisan language (Columns 1 and 3), or issue-polarizing language (Columns 2 and 4). Panel A shows the OLS estimates of β from Equation (1). Panel B shows the OLS estimates of β_1 , β_2 , and β_3 from Equation (2) with $a = b = c = 60$. Standard errors are two-way clustered at the event level and at the level of 30 minutes time intervals since the onset of the debate. Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

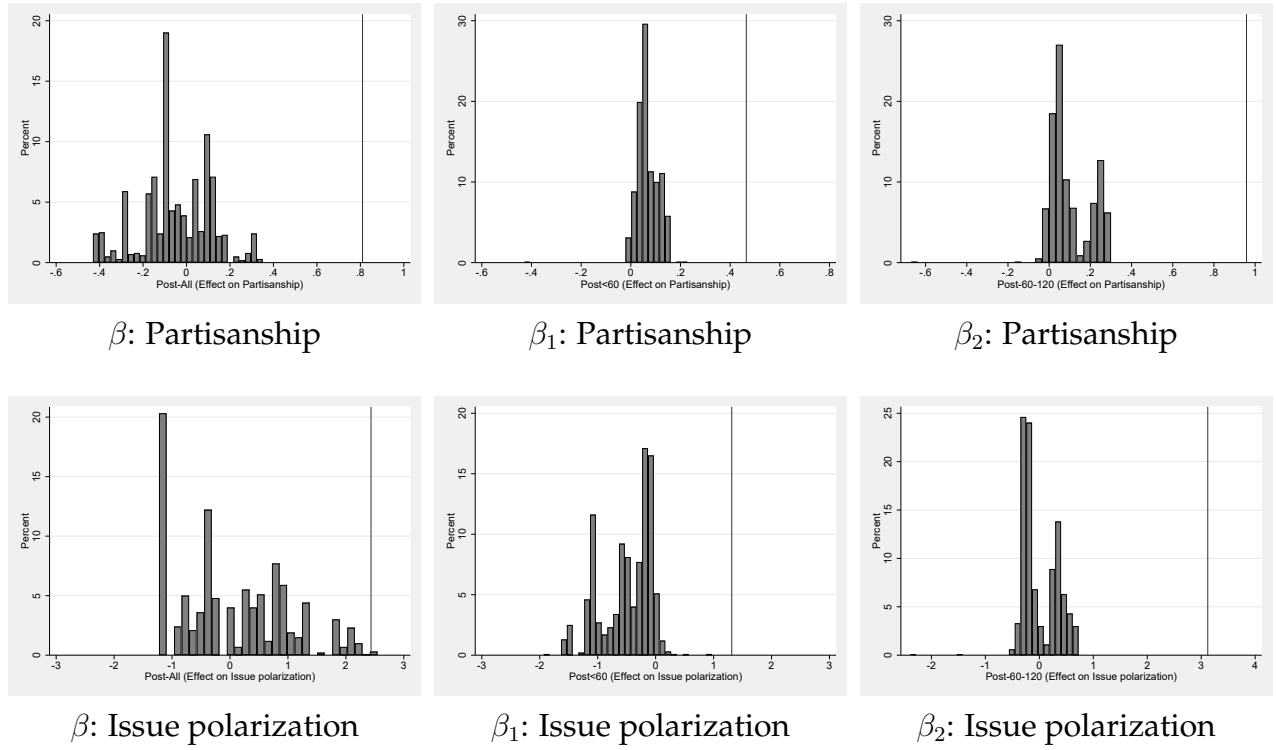
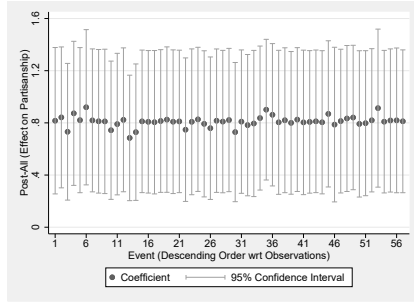
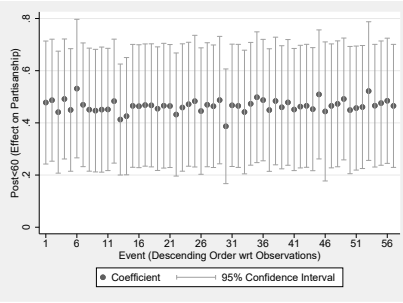


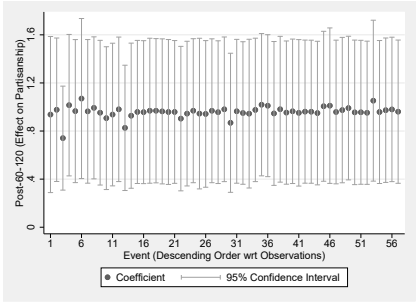
Figure A.8: First political intervention, partisanship and issue polarization: Randomization inference
This figure shows the distribution of β (left panels), β_1 (middle panels), and β_2 (right panels) from Equations (1) and (2) for Partisanship (top panel) and Issue polarization (bottom panel) where instead of using the real distribution of political interventions, we randomly reallocate the same number of political interventions across events. The results of this permutation inference with placebo treatments are based on 1,000 replications. The vertical bar indicates the coefficient obtained from the actual distribution of political interventions.



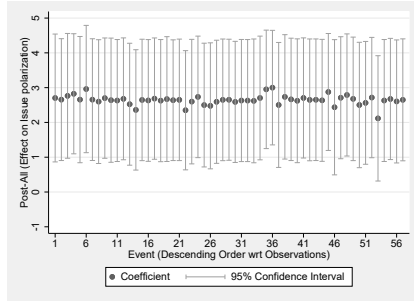
β for partisanship



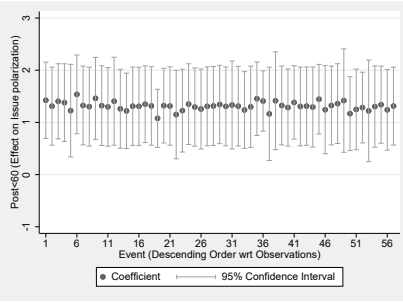
β_1 for partisanship



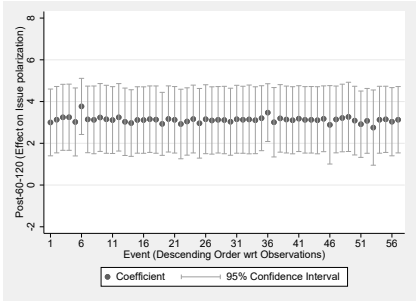
β_2 for partisanship



β for issue polarization



β_1 for issue polarization



β_2 for issue polarization

Figure A.9: Estimates from Equations (1) and (2), dropping one event at a time

The figure displays estimated β from Equation (1) (left panels) and estimated β_1 and β_2 from Equation (2) (middle and right panels), dropping one event at a time from the estimation sample.

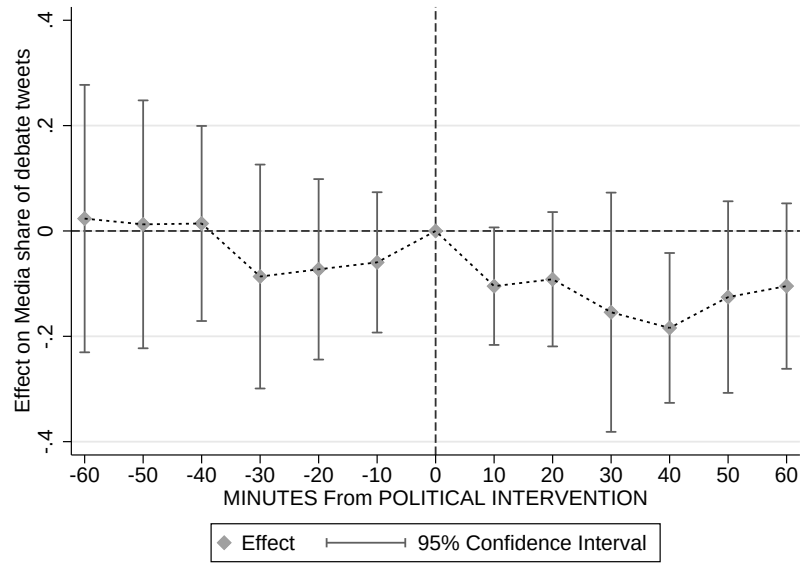


Figure A.10: News media activity around the first political intervention: Event-study estimates

The figure plots OLS coefficients with 95% confidence intervals from Equation (3) with the media share of the debate as the outcome, for the 60 minutes before and after the first political intervention. The unit of observation is a tweet and the dependent variable is equal to 100 (and zero otherwise) if the tweet is posted by one of the 76 news media organizations' accounts. The specification includes event fixed effects, fixed effects for 30-minute intervals since the onset of the debate, and a linear time trend. The omitted bin is the 10 minutes before the first political intervention. Standard errors are two-way clustered at the event level and at the level of 30-minute intervals since the onset of the debate.

Table A.13: First political intervention, partisanship and issue polarization: Robustness to news media activity

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Partisanship			Issue polarization		Media tweet	
POST-Intervention<1h (β_1)	0.465*** (0.120)	0.458*** (0.117)	0.464*** (0.120)	1.313*** (0.383)	1.187*** (0.375)	1.299*** (0.382)	-0.090 (0.069)
POST-Intervention 1-2h (β_2)	0.958*** (0.303)	0.952*** (0.307)	0.957*** (0.300)	3.122*** (0.811)	2.928*** (0.819)	3.105*** (0.808)	-0.126 (0.080)
POST-Intervention>2h (β_3)	0.983*** (0.360)	0.965** (0.364)	0.982*** (0.359)	4.229*** (1.369)	3.950*** (1.410)	4.215*** (1.372)	-0.147 (0.114)
Media tweets, previous 15 min		-0.007 (0.006)			-0.050** (0.020)		
Media tweets, previous 60 min		0.002 (0.004)			0.035** (0.014)		
Observations	4,747,621	4,747,621	4,714,492	4,747,621	4,747,621	4,714,492	4,747,621
R-squared	0.014	0.014	0.014	0.045	0.045	0.045	0.001
Media controls	No	Yes	No	No	Yes	No	No
Media tweets in sample	Included	Included	Dropped	Included	Included	Dropped	Included

The unit of observation is a tweet. The table shows OLS estimates of β_1 , β_2 , and β_3 from Equation (2) with $a = b = c = 60$. The dependent variable is equal to 100 (and zero otherwise) if the tweet contains partisan language (Columns 1–3), issue-polarizing language (Columns 4–6), or if the tweet is posted by a news media organization's account (Column 7; mean of 0.885 in the 60 minutes before the first political intervention). Columns 1 and 4 reproduce Panel B of Table 2. Columns 2 and 5 control for the number of tweets posted by news media accounts in the debate in the 15 and in the 60 minutes before the tweet. Columns 3 and 6 exclude tweets posted by news media accounts. In Column 7, the coefficient on tweets posted more than 60 minutes before the first political intervention (α) is 0.042 (standard error 0.057). Standard errors are two-way clustered at the event level and at the level of 30 minutes time intervals since the onset of the debate. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

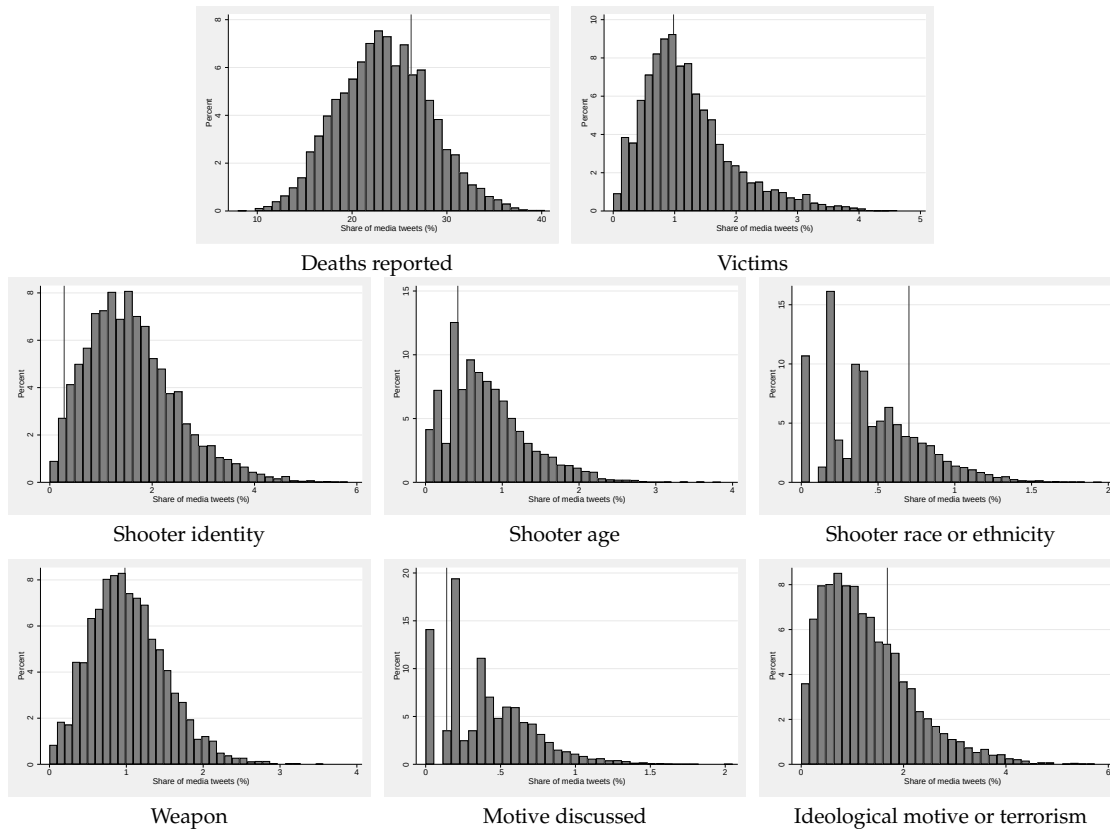
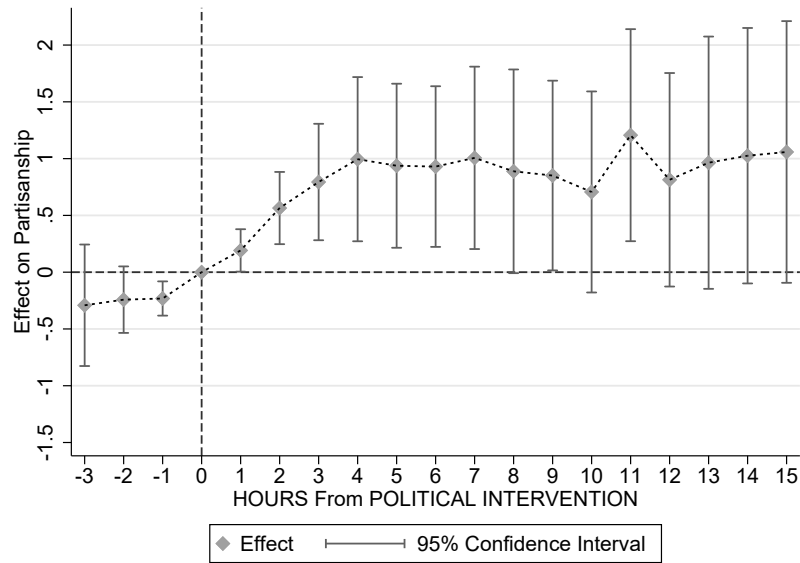


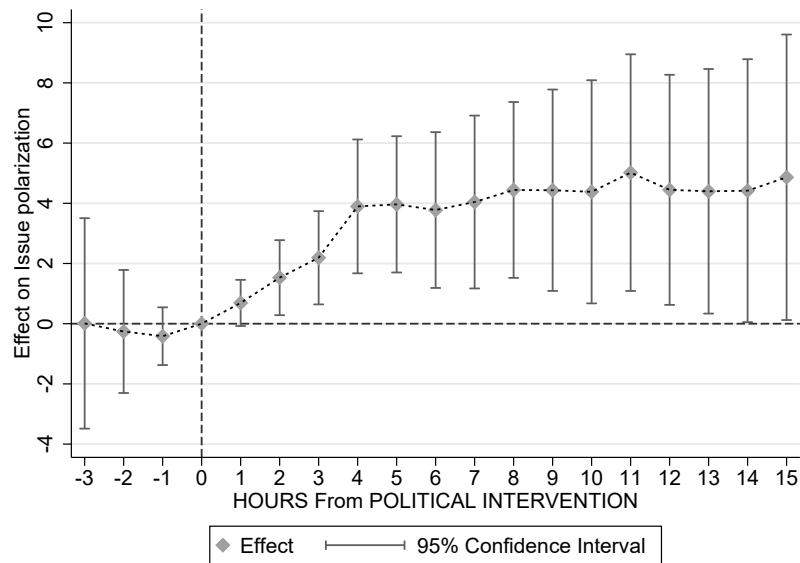
Figure A.11: Content of news media tweets before the first political intervention: Placebo test

Each panel refers to one content category. The vertical bar shows the share of news media tweets posted in the 30 minutes before each debate's first political intervention that contain the category, pooled over the 52 debates with a political intervention (713 news media tweets). The histogram shows the same share before placebo intervention times, across 5,000 placebo draws. In each draw, each debate's placebo time is drawn with replacement from the 52 actual first-intervention times (in minutes since the onset of the debate), so that placebo and actual interventions follow the same timing distribution; the timing and content of news media tweets are held fixed. A vertical bar inside the histogram indicates that news media content right before the first political intervention is not unusual relative to placebo times. Categories are identified with keyword-based rules applied to the tweet text. *Deaths reported*: a number (or "several," "multiple") followed by a word such as "killed" or "dead," or the phrase "death toll." *Victims*: who the victims were (e.g., "victims identified," "children among the dead"). *Shooter identity*: who the shooter is, by name or role (e.g., "identified as," "suspect named," "former employee"). *Shooter age*: e.g., "22-year-old gunman." *Shooter race or ethnicity*: e.g., "white man," "Hispanic suspect." *Weapon*: firearm details (e.g., "AR-15," "rifle," "semi-automatic," "high-capacity magazines"). *Motive discussed*: any mention of a motive (including "motive unknown"), a manifesto, a grievance, or a domestic or workplace dispute. *Ideological motive or terrorism*: e.g., "terrorism," "hate crime," "white supremacist," "ISIS."

A.5 Additional outcomes and long-term effects



(a) Partisanship



(b) Issue polarization

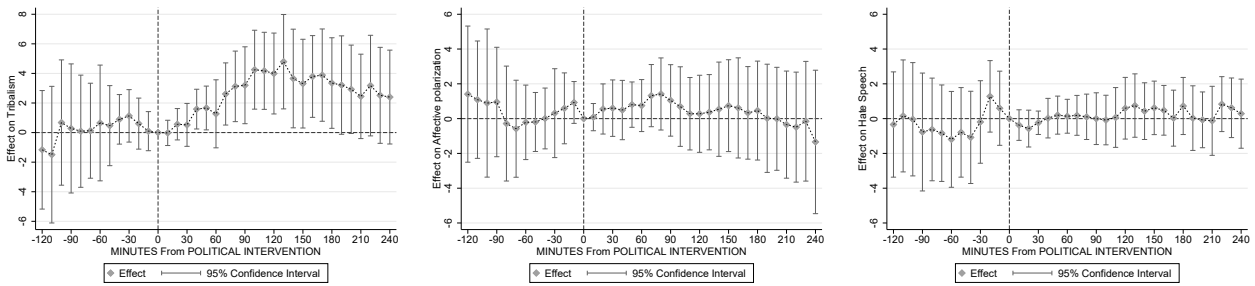
Figure A.12: Long-term event-study results

These figures plot OLS coefficients with 95% confidence intervals. The plotted coefficients are the event-studies coefficients associated with 1 hour windows, for a longer time horizon than our baseline specification. The outcome variable is Partisanship in panel (a) and Issue Polarization in panel (b). Standard errors are two-way clustered at the event level and at the level of 30 minutes time intervals since the onset of the debate.

Table A.14: First political intervention, tribalism, affective polarization, and hate speech

VARIABLES	(1) Tribalism	(2) Affective polarization	(3) Hate Speech
POST-Intervention (β)	1.601 (1.065)	-0.360 (1.559)	-0.040 (0.491)
Observations	4,747,621	4,747,621	4,747,621
R-squared	0.025	0.031	0.563
Time Trend	Linear	Linear	Linear
Equation	1	1	1
POST-Intervention<1h (β_1)	0.564 (0.803)	-0.011 (1.426)	-0.221 (0.356)
POST-Intervention 1-2h (β_2)	3.203*** (1.151)	0.432 (1.594)	0.078 (0.681)
POST-Intervention>2h (β_3)	2.271 (1.501)	-1.950 (2.329)	0.241 (0.856)
Observations	4,747,621	4,747,621	4,747,621
R-squared	0.025	0.031	0.563
Time Trend	Linear	Linear	Linear
Equation	2	2	2
Mean DV	8.90	37.71	12.58

The unit of observation is a tweet. The dependent variable is equal to 100 (and zero otherwise) if the tweet contains tribal language (column 1), or affective-polarized language (column 2), or hateful language (column 3). Panel A shows the OLS estimates of β from Equation (1). Panel B shows the OLS estimates of β_1 , β_2 , and β_3 from Equation (2) with $a = b = c = 60$. Standard errors are two-way clustered at the event level and at the level of 30 minutes time intervals since the onset of the debate. Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.



Tribalism

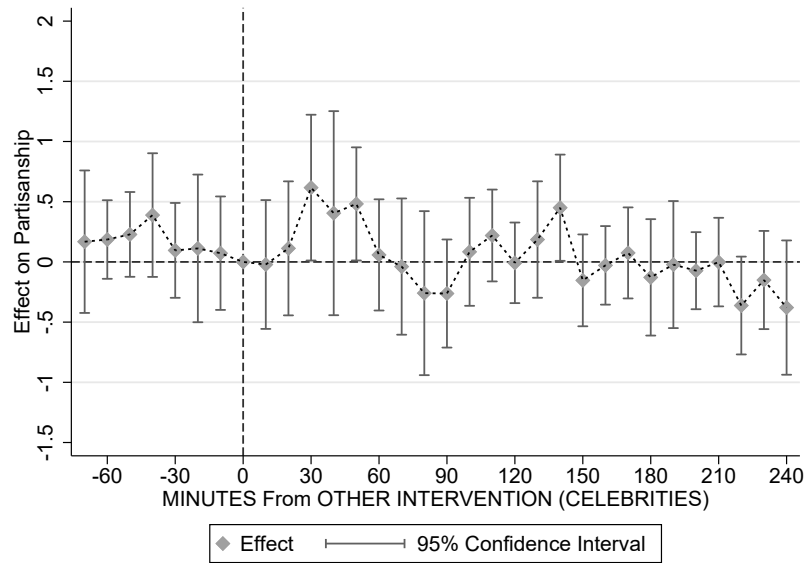
Affective Polarization

Hate Speech

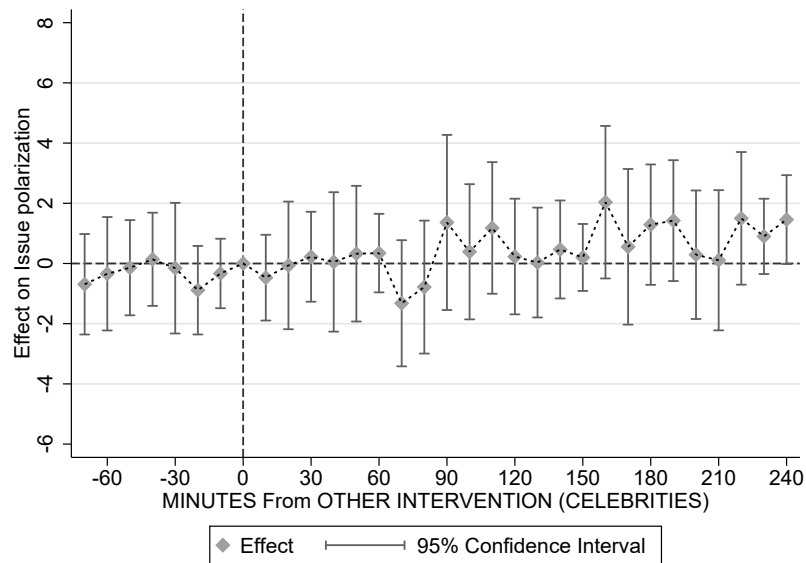
Figure A.13: Event-study results for tribalism, affective polarization, and hate speech

The figure plots OLS coefficients with 95% confidence intervals. The plotted coefficients are the β_τ coefficients associated with each 10 minute window from Equation (3). The outcome variable is tribalism, affective polarization, and hate speech in panels (a), (b), and (c), respectively. Standard errors are two-way clustered at the event level and at the level of 30 minutes time intervals since the onset of the debate.

A.6 Interventions by celebrities



(a) Partisanship



(b) Issue polarization

Figure A.14: Event-study results for other interventions (celebrities)

These figures plot OLS coefficients with 95% confidence intervals. The plotted coefficients are the β_τ coefficients associated with each 10 minute window, described in Equation (3) when p_e indicates the time when a celebrity tweets. The outcome variable is Partisanship in the top panel and Issue Polarization in the bottom panel. Standard errors are two-way clustered at the event level and at the level of 30 minutes time intervals since the onset of the debate.

Table A.15: First celebrity intervention, tribalism, affective polarization, and hate speech

VARIABLES	(1) Tribalism	(2) Affective polarization	(3) Hate Speech
POST-CELEBRITY (β)	-2.732** (1.171)	-5.432*** (1.439)	-0.396 (0.824)
Observations	4,748,501	4,748,501	4,748,501
R-squared	0.025	0.032	0.563
Time Trend	Linear	Linear	Linear
Equation	1	1	1
POST-CELEBRITY<1h (β_1)	0.749 (0.562)	0.018 (1.050)	0.449 (0.372)
POST-CELEBRITY 1-2h (β_2)	1.899* (1.068)	-1.896 (1.761)	1.386** (0.544)
POST-CELEBRITY>2h (β_3)	-3.588*** (1.217)	-6.394*** (1.566)	-0.670 (0.925)
Observations	4,748,501	4,748,501	4,748,501
R-squared	0.026	0.032	0.563
Time Trend	Linear	Linear	Linear
Equation	2	2	2
Mean DV	19.16	48.96	31.01

The unit of observation is a tweet. The dependent variable is equal to 100 (and zero otherwise) if the tweet contains tribal language (column 1), or affective-polarized language (column 2), or hateful language (column 3). Panel A shows the OLS estimates of β from Equation (1). Panel B shows the OLS estimates of β_1 , β_2 , and β_3 from Equation (2) with $a = b = c = 60$. Regressions use all tweets not posted by celebrities; politicians' tweets remain in this sample. Standard errors are two-way clustered at the event level and at the level of 30 minutes time intervals since the onset of the debate. Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

A.7 Rhetorical supply and heterogeneous responses

Table A.16: Politicians' rhetoric and partisanship and issue polarization of the public debate

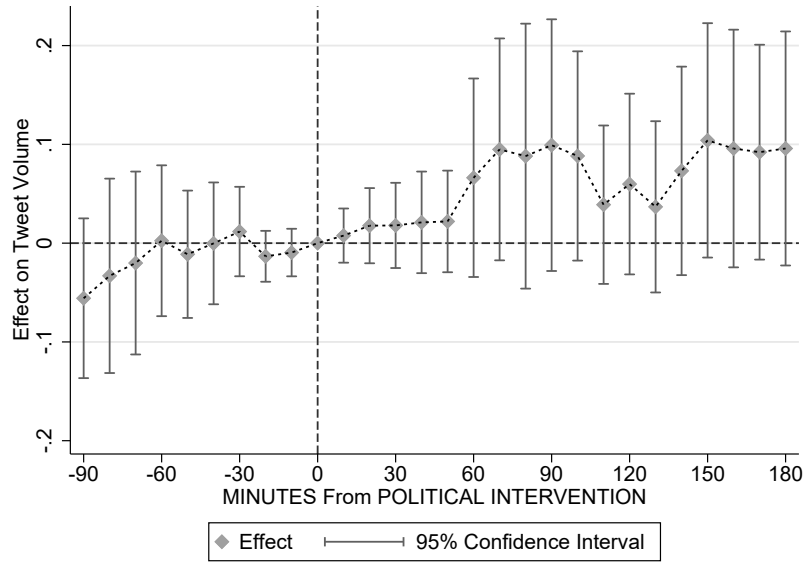
VARIABLES	(1) Partisanship	(2) Issue polarization	(3) Partisanship	(4) Issue polarization
SUPPLY=0 * POST-Intervention<1h	0.423** (0.182)	1.025* (0.540)	0.798*** (0.286)	2.236*** (0.802)
SUPPLY=0 * POST-Intervention 1-2h	0.974*** (0.353)	3.165*** (0.917)	0.897** (0.358)	3.756*** (1.155)
SUPPLY=0 * POST-Intervention>2h	0.957** (0.415)	3.971** (1.670)	1.059** (0.414)	4.689*** (1.644)
SUPPLY=1 * POST-Intervention<1h	1.525*** (0.545)	4.959** (1.991)	0.176 (0.332)	-0.025 (1.069)
SUPPLY=1 * POST-Intervention 1-2h	1.582*** (0.482)	4.200 (2.740)	1.437** (0.683)	2.486* (1.475)
SUPPLY=1 * POST-Intervention>2h	1.832*** (0.552)	7.508*** (1.822)	1.187* (0.693)	3.833* (2.145)
Observations	4,747,621	4,747,621	4,747,621	4,747,621
R-squared	0.014	0.045	0.014	0.045
Supply	Partisanship		Issue polarization	
Difference in <1h	1.102** (0.537)	3.934** (1.948)	-0.622 (0.463)	-2.261 (1.628)
Difference in 1-2h	0.608 (0.460)	1.035 (2.669)	0.540 (0.726)	-1.270 (1.927)
Difference in >2h	0.875 (0.542)	3.536* (2.099)	0.128 (0.627)	-0.857 (2.123)

The unit of observation is a tweet. The dependent variable is equal to 100 (and zero otherwise) if the tweet contains partisan language (columns 1 and 3), or issue-polarizing language (columns 2 and 4). Panel A shows the results of β_S and β_{NS} from OLS estimation of Equation 5 adapted to fit Equation (2). In columns 1 and 2, S captures whether the political intervention contains partisan language. In columns 3 and 4, S captures whether the political intervention contains issue-polarizing language. Standard errors are two-way clustered at the event level and at the level of 30 minutes time intervals since the onset of the debate. *** p<0.01, ** p<0.05, * p<0.1.

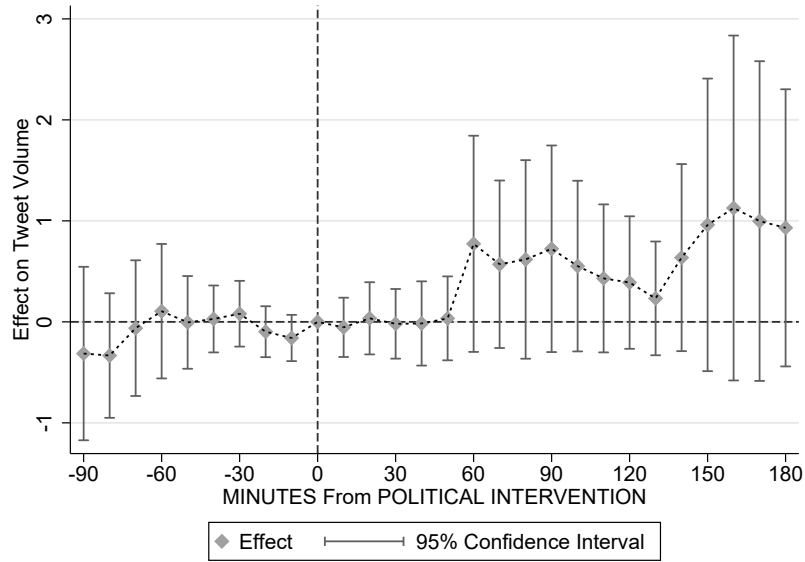
Table A.17: Politicians' characteristics and partisanship and issue polarization of the public debate in different post-estimation windows

VARIABLES	(1) Partisanship	(2) Issue polarization	(3) Partisanship	(4) Issue polarization
Type=0 * POST-Intervention<1h	0.475* (0.238)	1.523* (0.861)	0.928*** (0.284)	3.297*** (0.987)
Type=0 * POST-Intervention 1-2h	0.979** (0.450)	2.385* (1.209)	1.137*** (0.337)	5.400*** (1.355)
Type=0 * POST-Intervention>2h	0.986** (0.441)	3.553* (1.806)	0.959** (0.464)	4.002** (1.975)
Type=1 * POST-Intervention<1h	0.957*** (0.279)	3.887*** (1.362)	0.459** (0.227)	1.007 (0.624)
Type=1 * POST-Intervention 1-2h	1.436*** (0.280)	7.283*** (1.493)	1.045** (0.397)	2.834*** (0.986)
Type=1 * POST-Intervention>2h	1.439*** (0.411)	8.009*** (2.276)	1.090** (0.441)	4.400*** (1.587)
Observations	4,747,621	4,747,621	4,747,621	4,747,621
R-squared	0.014	0.045	0.014	0.045
Type	Republican		Male	
Difference in <1h	0.482 (0.366)	2.364 (1.533)	-0.470 (0.317)	-2.290** (1.084)
Difference in 1-2h	0.457 (0.408)	4.924*** (1.673)	-0.093 (0.406)	-2.566* (1.414)
Difference in >2h	0.454 (0.307)	4.456** (1.962)	0.131 (0.409)	0.398 (1.803)

The unit of observation is a tweet. The dependent variable is equal to 100 (and zero otherwise) if the tweet contains partisan language (columns 1 and 3), or issue-polarizing language (columns 2 and 4). Panel A shows the results of β_S and β_{NS} from OLS estimation of Equation 5 adapted to fit Equation (2). In columns 1 and 2, S captures whether the political intervention is from a Republican. In columns 3 and 4, S captures whether the political intervention is from a male politician. Standard errors are two-way clustered at the event level and at the level of 30 minutes time intervals since the onset of the debate. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.



(a) Politicized users



(b) Non-politicized users

Figure A.15: Volume of tweets for politicized and non-politicized Users: Event-study results

These figures plot OLS coefficients with 95% confidence intervals. The plotted coefficients are the β_τ coefficients associated with each 10 minute window from Equation (3). The outcome variable is Number of Tweets by Politicized Users in panel (a) and Non-Politicized Users in panel (b). Standard errors are two-way clustered at the event level and at the level of 30 minutes time intervals since the onset of the debate.

Table A.18: Politicians' characteristics and partisan language in the public debate: Democratic-leaning, Republican-leaning, and mixed partisan language

VARIABLES	(1) Democratic-leaning phrases only	(2) Republican-leaning phrases only	(3) Both Democratic- and Republican-leaning phrases
Party=0 * POST-Intervention	0.505*** (0.165)	0.215 (0.214)	-0.004 (0.025)
Party=1 * POST-Intervention	0.565*** (0.153)	0.602*** (0.203)	0.021 (0.019)
Observations	4,747,621	4,747,621	4,747,621
R-squared	0.005	0.021	0.001
Party		Republican	
Difference	0.060	0.387**	0.025
SE	(0.207)	(0.189)	(0.015)

The unit of observation is a tweet. The dependent variable is equal to 100 (and zero otherwise) if the tweet contains Democratic-leaning partisan phrases only (column 1), Republican-leaning partisan phrases only (column 2), a Democratic-leaning partisan phrase and a Republican-leaning partisan phrase, which we refer to as bi-partisan (column 3). Panel A shows the results of β_S and β_{NS} from OLS estimation of Equation 5. In columns 1, 2, and 3, S captures whether the political intervention is from a Republican. Standard errors are two-way clustered at the event level and at the level of 30 minutes time intervals since the onset of the debate. Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.19: Politicians' characteristics and partisan language in the public debate: Democratic-leaning, Republican-leaning, and mixed partisan language in different post-estimation windows

VARIABLES	(1) Democratic-leaning phrases only	(2) Republican-leaning phrases only	(3) Both Democratic- and Republican-leaning phrases
Party=0 * POST-Intervention<1h	0.485*** (0.144)	-0.018 (0.122)	0.008 (0.020)
Party=0 * POST-Intervention 1-2h	0.592*** (0.181)	0.363 (0.356)	0.025 (0.032)
Party=0 * POST-Intervention>2h	0.577** (0.271)	0.395 (0.323)	0.014 (0.035)
Party=1 * POST-Intervention<1h	0.379** (0.145)	0.606** (0.266)	-0.028 (0.023)
Party=1 * POST-Intervention 1-2h	0.733*** (0.197)	0.691*** (0.232)	0.013 (0.025)
Party=1 * POST-Intervention>2h	0.639** (0.240)	0.758** (0.288)	0.042 (0.030)
Observations	4,747,621	4,747,621	4,747,621
R-squared	0.005	0.021	0.001
Party		Republican	
Difference in <1h	-0.106 (0.203)	0.623** (0.302)	-0.036 (0.025)
Difference in 1-2h	0.140 (0.237)	0.328 (0.327)	-0.012 (0.033)
Difference in >2h	0.061 (0.228)	0.364* (0.195)	0.029 (0.018)

The unit of observation is a tweet. The dependent variable is equal to 100 (and zero otherwise) if the tweet contains Democratic-leaning partisan phrases only (column 1), Republican-leaning partisan phrases only (column 2), a Democratic-leaning partisan phrase and a Republican-leaning partisan phrase, which we refer to as bi-partisan (column 3). Panel A shows the results of β_S and β_{NS} from OLS estimation of Equation 5 adapted to fit Equation (2). In columns 1, 2 and 3, S captures whether the first political intervention is from a Republican. Standard errors are two-way clustered at the event level and at the level of 30 minutes time intervals since the onset of the debate. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.20: Responses to political interventions by returning vs. new users and politicized vs. non politicized users

VARIABLES	(1) Partisanship	(2) Issue polarization	(3) Partisanship	(4) Issue polarization
POST-Intervention (β) $\times$ (User Type=0)	1.023*** (0.281)	3.885*** (0.846)	0.772*** (0.276)	1.758* (0.901)
POST-Intervention (β) $\times$ (User Type=1)	0.148 (0.262)	-1.213 (1.325)	1.200*** (0.339)	11.651*** (1.155)
Observations	4,747,621	4,747,621	4,747,621	4,747,621
R-squared	0.014	0.047	0.014	0.051
Time Trend	Linear	Linear	Linear	Linear
Equation	1	1	1	1
User Type	Returning	Returning	Politicized	Politicized
Mean DV	1.08	7.75	1.08	7.75
Type=1 - Type=0	-0.874*** (0.110)	-5.099*** (1.009)	0.428** (0.186)	9.893*** (0.599)

The unit of observation is a tweet. The dependent variable is equal to 100 (and zero otherwise) if the tweet contains partisan language (columns 1 and 3), or issue-polarizing language (columns 2 and 4). Panel A shows the results of β_S and β_{NS} from OLS estimation of Equation 5. In columns 1 and 2, the user was a returning user (user who already posted before the political intervention). In columns 3 and 4, the user was a political user (user with political text in account description). Standard errors are two-way clustered at the event level and at the level of 30 minutes time intervals since the onset of the debate. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

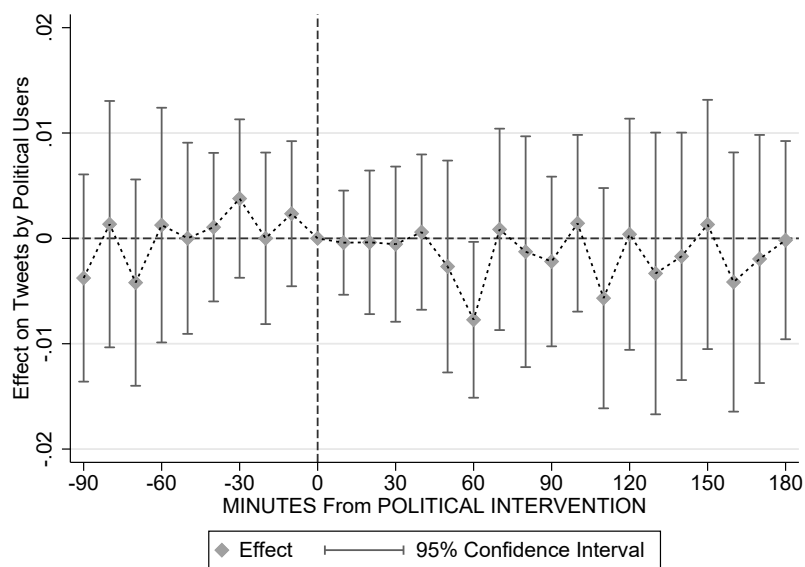


Figure A.16: Share of politicized users: Event-study results

This figure plots OLS coefficients with 95% confidence intervals. The plotted coefficients are the β_τ coefficients associated with each 10 minute window from Equation (3). The outcome variable is the proportion of Tweets by Politicized Users (of total tweets). Standard errors are two-way clustered at the event level and at the level of 30 minutes time intervals since the onset of the debate.